



IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re U.S. Patent Application of)
AIZAWA et al.) Art Unit 2655
Application Number: 10/782,925) Examiner Dieu Minh T. LE
Filed: February 23, 2004)
For: DISK ARRAY SYSTEM AND A METHOD OF)
AVOIDING FAILURE OF THE DISK ARRAY)
SYSTEM)
Attorney Docket No. HITA.0522)

**Honorable Assistant Commissioner
for Patents
Washington, D.C. 20231**

LETTER

Sir:

The below-identified communications are submitted in the above-captioned application or proceeding:

- (x) Request for Priority
- (x) Certified copy of Japanese Patent Application 2003-395322
- (x) Copy of Request for Priority filed July 22, 2004
- (x) Postcard Stamped by USPTO Mailroom

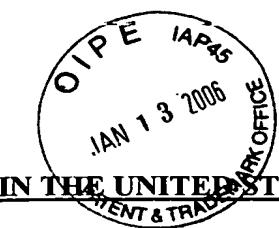
The Commissioner is hereby authorized to charge payment of any fees associated with this communication, including fees under 37 C.F.R. § 1.16 and 1.17 or credit any overpayment to Deposit Account Number 08-1480. A duplicate copy of this sheet is attached.

Respectfully submitted,

Stanley P. Fisher
Registration Number 24,344

Juan Carlos A. Marquez
Registration Number 34,072

REED SMITH LLP
3110 Fairview Park Drive
Suite 1400
Falls Church, Virginia 22042
(703) 641-4200
January 13, 2006



IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re U.S. Patent Application of)
AIZAWA et al.) Art Unit 2655
Application Number: 10/782,925) Examiner Dieu Minh T. LE
Filed: February 23, 2004)
For: DISK ARRAY SYSTEM AND A METHOD OF)
AVOIDING FAILURE OF THE DISK ARRAY)
SYSTEM)
Attorney Docket No. HITA.0522)

Honorable Assistant Commissioner
for Patents
Washington, D.C. 20231

**REQUEST FOR PRIORITY
UNDER 35 U.S.C. § 119
AND THE INTERNATIONAL CONVENTION**

Sir:

In the matter of the above-captioned application for a United States patent, notice is hereby given that the Applicant claims the priority date of November 26, 2003, the filing date of the corresponding Japanese patent application 2003-395322

A certified copy of Japanese patent application 2003-395322 was submitted on July 22, 2004. Acknowledgment of receipt of the certified copy was respectfully requested, but no such receipt has been forthcoming. In a telephone conference with the Examiner on November 2, 2005, the Examiner stated that the priority document is missing and therefore in order to receive the acknowledgement of the priority claim, Applicant must submit a certified copy of Japanese patent application 2003-395322. Also submitted herewith is the certified copy together with a copy of the Request for Priority submitted July 22, 2004 and postcard stamped by the USPTO mailroom.

Applicant respectfully requests acknowledgement of the priority in a Supplemental Notice of Allowability.

Respectfully submitted,

Stanley P. Fisher
Registration Number 24,344

Juan Carlos A. Marquez
Registration Number 34,072

REED SMITH LLP
3110 Fairview Park Drive
Suite 1400
Falls Church, Virginia 22042
(703) 641-4200
January 13, 2006

日本国特許庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日 Date of Application: 2003年11月26日

出願番号 Application Number: 特願2003-395322

パリ条約による外国への出願に用いる優先権の主張の基礎となる出願の国コードと出願番号

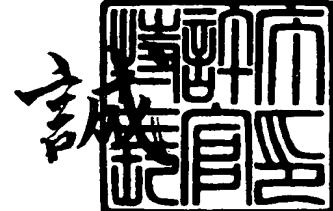
The country code and number of your priority application, to be used for filing abroad under the Paris Convention, is

出願人 Applicant(s): 株式会社日立製作所

特許庁長官
Commissioner,
Japan Patent Office

2005年12月5日

中 鳴



出証番号 出証特2005-3099873

【書類名】 特許願
【整理番号】 340301312
【あて先】 特許庁長官殿
【国際特許分類】 G06F 03/06
【発明者】
【住所又は居所】 神奈川県小田原市中里322番2号 株式会社日立製作所 R A I
【氏名】 相澤 正樹
【発明者】
【住所又は居所】 神奈川県小田原市中里322番2号 株式会社日立製作所 R A I
【氏名】 葛城 栄寿
【発明者】
【住所又は居所】 神奈川県小田原市中里322番2号 株式会社日立製作所 R A I
【氏名】 福岡 幹夫
【発明者】
【住所又は居所】 神奈川県小田原市中里322番2号 株式会社日立製作所 R A I
【氏名】 岡本 岳樹
【特許出願人】
【識別番号】 000005108
【氏名又は名称】 株式会社日立製作所
【代理人】
【識別番号】 100095371
【弁理士】
【氏名又は名称】 上村 輝之
【選任した代理人】
【識別番号】 100089277
【弁理士】
【氏名又は名称】 宮川 長夫
【選任した代理人】
【識別番号】 100104891
【弁理士】
【氏名又は名称】 中村 猛
【手数料の表示】
【予納台帳番号】 043557
【納付金額】 21,000円
【提出物件の目録】
【物件名】 特許請求の範囲 1
【物件名】 明細書 1
【物件名】 図面 1
【物件名】 要約書 1
【包括委任状番号】 0110323

【書類名】特許請求の範囲**【請求項 1】**

上位装置とのデータ授受を制御するチャネルアダプタと、
 R A I D グループを構成する複数のデータディスクドライブと、
 前記各データディスクドライブの予備として少なくとも 1 つ設けられる予備ディスクドライブと、
 前記各データディスクドライブ及び前記予備ディスクドライブとのデータ授受を制御するディスクアダプタと、
 前記チャネルアダプタ及び前記ディスクアダプタにより使用され、データを記憶するキッシュメモリと、
 前記チャネルアダプタ及び前記ディスクアダプタにより使用され、制御情報を記憶する制御メモリと、
 前記各データディスクドライブ及び前記予備ディスクドライブとは別に設けられる退避用記憶部と、
 前記ディスクアダプタに設けられ、前記各データディスクドライブに対するアクセスエラーの発生を監視して前記アクセスエラーの発生頻度が予め設定された所定の閾値以上になった場合には、前記閾値以上のデータディスクドライブに記憶されたデータを前記キッシュメモリを介して前記予備ディスクドライブにコピーさせる第 1 制御部と、
 前記ディスクアダプタに設けられ、前記第 1 制御部による前記コピー中に前記 R A I D グループを対象とするアクセス要求を処理し、前記 R A I D グループを対象とする書き込み要求を前記退避用記憶部に対して実行させる第 2 制御部と、
 前記ディスクアダプタに設けられ、前記第 1 制御部による前記コピーが終了した場合に前記第 2 制御部により前記退避用記憶部に書き込まれたデータを、前記閾値以上のデータディスクドライブ以外の前記各データディスクドライブ及び前記予備ディスクドライブに反映させる第 3 制御部と、を含んで構成されるディスクアレイ装置。

【請求項 2】

前記第 2 制御部は、前記閾値以上のデータディスクドライブを対象とする読み出し要求を、前記閾値以上のデータディスクドライブ以外の前記各データディスクドライブ内に記憶されたデータに基づいて処理する請求項 1 に記載のディスクアレイ装置。

【請求項 3】

前記第 2 制御部は、前記閾値以上のデータディスクドライブ以外の前記各データディスクドライブを対象とする読み出し要求を、前記退避用記憶部にコピーされたデータに基づいて処理する請求項 1 に記載のディスクアレイ装置。

【請求項 4】

前記第 2 制御部は、前記退避用記憶部に書き込まれたデータを管理する差分管理情報に関連付けられており、この差分管理情報に基づいて、前記 R A I D グループを対象とする読み出し要求を、前記閾値以上のデータディスクドライブ以外の前記各データディスクドライブ内に記憶されたデータに基づいて処理するか、あるいは前記退避用記憶部に記憶されたデータに基づいて処理するかを決定する請求項 1 に記載のディスクアレイ装置。

【請求項 5】

前記第 2 制御部は、前記 R A I D グループを対象とする書き込み要求のうち前記閾値以上のデータディスクドライブへの書き込み要求のみを前記退避用記憶部に対して実行させ、前記閾値以上のデータディスクドライブ以外の前記各データディスクドライブへの書き込み要求は、当該各データディスクドライブに対して実行させる請求項 1 に記載のディスクアレイ装置。

【請求項 6】

前記第 2 制御部は、前記退避用記憶部に所定値以上の空き容量がある場合に、前記 R A I D グループを対象とする書き込み要求を前記退避用記憶部に対して実行させ、前記退避用記憶部に前記所定値以上の空き容量が無い場合に、前記 R A I D グループを対象とする書き込み要求を前記 R A I D グループに対して実行させる請求項 1 に記載のディスクアレイ装

置。

【請求項 7】

前記第1制御部は、前記閾値以上のデータディスクドライブ以外の前記各データディスクドライブ内に記憶されたデータに基づいて、前記閾値以上のデータディスクドライブ内のデータを復元し、この復元されたデータを前記予備ディスクドライブにコピーさせるものである請求項1に記載のディスクアレイ装置。

【請求項 8】

前記第1制御部によるコピー処理実行させる手動指示部を設けた請求項1に記載のディスクアレイ装置。

【請求項 9】

前記第1制御部及び前記第2制御部は多重動作可能であり、前記退避用記憶部は、複数のRAIDグループのそれぞれを対象とする書き込み要求を受け入れるようになっている請求項1に記載のディスクアレイ装置。

【請求項 10】

前記退避用記憶部は、少なくとも、前記RAIDグループと同一構成を有する別のRAIDグループ、論理ボリューム、ディスクドライブのいずれか1つとして実現される請求項1に記載のディスクアレイ装置。

【請求項 11】

上位装置とのデータ授受を制御するチャネルアダプタと、RAIDグループを構成する複数のデータディスクドライブと、前記各データディスクドライブの予備として少なくとも1つ設けられる予備ディスクドライブと、前記各データディスクドライブ及び前記予備ディスクドライブとのデータ授受を制御するディスクアダプタと、前記チャネルアダプタ及び前記ディスクアダプタにより使用され、データを記憶するキャッシングメモリと、前記チャネルアダプタ及び前記ディスクアダプタにより使用され、制御情報を記憶する制御メモリと、前記各データディスクドライブ及び前記予備ディスクドライブとは別に設げられる退避用記憶部と、を含んだディスクアレイ装置の障害回避方法であって、

前記各データディスクドライブに対するアクセスエラーの発生を監視し、前記アクセスエラーの発生頻度が予め設定された所定の閾値以上になったか否かを判定する第1ステップと、

前記第1ステップにより前記閾値以上のデータディスクドライブが検出された場合は、この閾値以上のデータディスクドライブに記憶されたデータを前記予備ディスクドライブにコピーさせる第2ステップと、

前記第1ステップによる前記コピーの開始によって、前記RAIDグループと前記退避用記憶部とを関連付ける第3ステップと、

前記第1ステップによる前記コピー中に、前記RAIDグループを対象とするアクセス要求が発生したか否かを判定する第4ステップと、

前記第4ステップにより前記アクセス要求の発生が検出された場合、前記アクセス要求が書き込み要求であるならば、前記第3ステップにより関連付けられた前記退避用記憶部に対してデータを書き込む第5ステップと、

を含むディスクアレイ装置の障害回避方法。

【請求項 12】

前記第2ステップによる前記コピーが終了した場合に、前記5ステップにより前記退避用記憶部に書き込まれたデータを、前記閾値以上のデータディスクドライブ以外の前記各データディスクドライブ及び前記予備ディスクドライブに反映させる第6ステップをさらに含んだ請求項11に記載のディスクアレイ装置の障害回避方法。

【請求項 13】

前記第5ステップは、前記第4ステップにより検出された前記アクセス要求が前記閾値以上のデータディスクドライブを対象とする読み出し要求であるならば、この読み出し要求を、前記閾値以上のデータディスクドライブ以外の前記各データディスクドライブ内に記憶されたデータに基づいて処理するようになっている請求項11に記載のディスクアレイ裝

置の障害回避方法。

【請求項14】

前記第5ステップは、前記退避用記憶部に書き込まれたデータを管理する差分管理情報を利用することにより、前記第4ステップにより検出された前記RAIDグループを対象とする読み出し要求を、前記閾値以上のデータディスクドライブ以外の前記各データディスクドライブ内に記憶されたデータに基づいて処理するか、あるいは前記退避用記憶部に記憶されたデータに基づいて処理するかを決定する請求項11に記載のディスクアレイ装置の障害回避方法。

【請求項15】

前記第5ステップは、前記第4ステップにより検出された前記RAIDグループを対象とする書き込み要求のうち、前記閾値以上のデータディスクドライブへの書き込み要求のみを前記退避用記憶部に対して実行させ、前記閾値以上のデータディスクドライブ以外の前記各データディスクドライブへの書き込み要求は当該各データディスクドライブに対して実行させる請求項11に記載のディスクアレイ装置の障害回避方法。

【請求項16】

前記第2ステップは、前記閾値以上のデータディスクドライブ以外の前記各データディスクドライブ内に記憶されたデータに基づいて、前記閾値以上のデータディスクドライブに記憶されているデータを復元し、この復元されたデータを前記予備ディスクドライブにコピーさせるものである請求項11に記載のディスクアレイ装置の障害回避方法。

【請求項17】

RAIDグループを構成する複数のディスクドライブを含んだディスクアレイ装置のディスクドライブ使用方法であって、

前記RAIDグループを構成する前記各ディスクドライブに対するアクセスエラーの発生を監視し、前記アクセスエラーの発生頻度が予め設定された所定の閾値以上になった場合に障害ディスクドライブであると判定する障害ドライブ検出ステップと、

前記障害ドライブ検出ステップによって前記障害ディスクドライブが検出された場合は、この障害ディスクドライブに記憶されたデータを、前記RAIDグループを構成する前記各ディスクドライブ以外の正常ディスクドライブにコピーさせるデータコピーステップと、

前記データコピーステップによる前記コピー中に、前記RAIDグループを対象とするアクセス要求が発生したか否かを検出するアクセス要求検出ステップと、

前記アクセス要求検出ステップにより書き込み要求が検出された場合は、前記データコピーがされている正常ディスクドライブとは別の正常ディスクドライブに対して、前記書き込み要求に係わるデータを書き込むアクセス処理ステップと、

を含むディスクアレイ装置のディスクドライブ使用方法。

【請求項18】

前記データコピーステップによる前記データコピーが終了した場合に、前記アクセス処理ステップにより前記正常ディスクドライブに書き込まれたデータを、前記障害ディスクドライブ以外の前記RAIDグループを構成する前記各ディスクドライブ及び前記データコピーされた正常ディスクドライブに反映させるデータ更新ステップをさらに含んだ請求項17に記載のディスクアレイ装置のディスクドライブ使用方法。

【請求項19】

前記アクセス処理ステップは、前記アクセス要求検出ステップによって前記障害ディスクドライブを対象とする読み出し要求が検出された場合、前記RAIDグループを構成する前記障害ディスクドライブ以外の前記各ディスクドライブ内に記憶されたデータに基づいて、要求されたデータを復元する請求項17に記載のディスクアレイ装置のディスクドライブ使用方法。

【請求項20】

前記データコピーステップは、前記RAIDグループを構成する前記障害ディスクドライブ以外の前記各ディスクドライブに記憶されたデータに基づいて、前記障害ディスクド

ライブに記憶されているデータを復元し、この復元されたデータを前記正常ディスクドライブにデータコピーさせる請求項17に記載のディスクアレイ装置のディスクドライブ使用方法。

【書類名】明細書

【発明の名称】ディスクアレイ装置及びディスクアレイ装置の障害回避方法

【技術分野】

【0001】

本発明は、複数のディスクドライブを有するディスクアレイ装置及びディスクアレイ装置の障害回避方法に関する。

【背景技術】

【0002】

ディスクアレイ装置は、例えば、多数のディスクドライブをアレイ状に配設し、RAID (Redundant

Array of Independent Inexpensive Disks) に基づいて構築されている。各ディスク装置が有する物理的な記憶領域上には、論理的な記憶領域である論理ボリュームが形成されている。ホストコンピュータは、ディスクアレイ装置に対して所定形式の書き込みコマンド又は読み出しコマンドを発行することにより、所望のデータの読み書きを行うことができる。

【0003】

ディスクアレイ装置には、ディスクドライブに記憶したデータの消失等を防止するために、種々の防御策が施されている。1つは、RAID構成の採用である。例えば、RAID 1～6 等として知られている冗長記憶構造をディスクアレイ装置が採用することにより、データ消失の可能性が低減される。これに加えて、ディスクアレイ装置では、例えば、RAID構造の論理ボリュームを二重化し、正ボリュームと副ボリュームの一対の論理ボリュームにそれぞれ同一のデータを記憶させることもできる。また、いわゆるディザスタリカバリとして知られているように、自然災害等の不測の事態に備えて、ローカルサイトから遠く離れたリモートサイトに、データのコピーを保存する場合もある。また、ディスクアレイ装置に記憶されているデータは、定期的に、テープドライブ等のバックアップ装置に記憶される。

【0004】

さらに、ディスクアレイ装置では、物理的構成の二重化も行われている。例えば、ディスクアレイ装置では、ホストコンピュータとの間のデータ通信を行う上位インターフェース回路や各ディスクドライブとの間のデータ通信を行う下位インターフェース回路等の主要部を複数設けて多重化している。また、これら各主要部間をそれぞれ接続する経路や、各主要部に電力を供給する電源等も複数設けられている。

【0005】

これらに加えて、ディスクアレイ装置は、予備のディスクドライブを1つ以上備えることができる。データが記憶されているディスクドライブに何らかの障害が発生した場合、この障害の発生したディスクドライブに記憶されているデータは、予備ディスクにコピーされる。例えば、他のディスクドライブに分散して記憶されているデータ及びパリティに基づいて逆演算することにより、障害の発生したディスクドライブ内のデータを復元する(特許文献1)。その後、障害の発生したディスクドライブを取り出し、新品のディスクドライブや予備ディスクドライブに入れ替える。

【特許文献1】特開平7-146760号公報

【発明の開示】

【発明が解決しようとする課題】

【0006】

従来技術では、ディスクドライブに障害が発生した場合に、障害ディスクドライブに記憶されているデータを、他の正常なディスクドライブ記憶されたデータとパリティとに基づいて復元する。そして、従来技術では、復元したデータを予備ディスクドライブに格納する。このように、従来技術では、あるディスクドライブに実際に障害が発生するまでは、予備ディスクドライブへのデータコピーが行われない。従って、予備ディスクドライブへのデータコピー開始時期に遅れが発生する。また、正常なディスクドライブからデータを復元するため、データ復元に時間がかかり、データコピー完了までの時間もかかる。

【0007】

さらに、続けて他の正常なディスクドライブの一部に何らかの障害が発生した場合は、逆演算に必要なデータを取得できないため、障害の発生したディスクドライブのデータを復元することができない。正常なディスクドライブであっても、読み書きを繰り返すことにより、部分的な障害を引き起こす可能性が増加する。2つ以上の情報（データ、パリティ）が読み出しが不能になった場合は、逆演算によりデータを復元できないため、復元不能なデータは失われることになる。

【0008】

本発明の1つの目的は、障害発生のおそれのあるディスクドライブから予備ディスクドライブへのデータ移行を従来よりも安全に行うことができるディスクアレイ装置及びディスクアレイ装置の障害回避方法を提供することにある。本発明の1つの目的は、障害発生のおそれのあるディスクドライブ以外の正常なディスクドライブへの読み書きを低減することにより、正常なディスクドライブに障害が発生する可能性を低減できるようにしたディスクアレイ装置及びディスクアレイ装置の障害回避方法を提供することにある。本発明の他の目的は、後述する実施の形態の記載から明らかになるであろう。

【課題を解決するための手段】**【0009】**

上記課題を解決すべく、本発明に従うディスクアレイ装置は、上位装置とのデータ授受を制御するチャネルアダプタと、RAIDグループを構成する複数のデータディスクドライブと、各データディスクドライブの予備として少なくとも1つ設けられる予備ディスクドライブと、各データディスクドライブ及び予備ディスクドライブとのデータ授受を制御するディスクアダプタと、チャネルアダプタ及びディスクアダプタにより使用され、データを記憶するキャッシュメモリと、チャネルアダプタ及びディスクアダプタにより使用され、制御情報を記憶する制御メモリと、各データディスクドライブ及び予備ディスクドライブとは別に設けられる退避用記憶部と、ディスクアダプタに設けられ、各データディスクドライブに対するアクセスエラーの発生を監視してアクセスエラーの発生頻度が予め設定された所定の閾値以上になった場合には、閾値以上のデータディスクドライブに記憶されたデータをキャッシュメモリを介して予備ディスクドライブにコピーさせる第1制御部と、ディスクアダプタに設けられ、第1制御部によるコピー中にRAIDグループを対象とするアクセス要求を処理し、RAIDグループを対象とする書き込み要求を退避用記憶部に対して実行させる第2制御部と、ディスクアダプタに設けられ、第1制御部によるコピーが終了した場合に第2制御部により退避用記憶部に書き込まれたデータを、閾値以上のデータディスクドライブ以外の各データディスクドライブ及び予備ディスクドライブに反映させる第3制御部と、を含んで構成されている。

【0010】

チャネルアダプタは、上位装置から受信したデータをキャッシュメモリに格納する。また、チャネルアダプタは、上位装置から受信したコマンド（読み出し命令、書き込み命令等）を制御メモリに格納する。ディスクアダプタは、制御メモリの内容を参照することにより、上位装置からの受信データをキャッシュメモリから読み出して、所定のデータディスクドライブに記憶させる（書き込み命令の場合）。また、ディスクアダプタは、制御メモリの内容を参照することにより、上位装置から要求されたデータをデータディスクドライブから読み出して、キャッシュメモリに格納させる（読み出し命令の場合）。チャネルアダプタは、キャッシュメモリに格納されたデータを読み出して上位装置に送信する。

さて、RAIDグループを構成する複数のデータディスクドライブには、データ（パリティを含む）が分散して記憶されている。例えば、RAID5では、パリティ専用のディスクドライブを備えておらず、通常のデータと同様に、パリティもデータディスクドライブに分散して記憶される。退避用記憶部は、RAIDグループに対する書き込み要求を処理するために設けられており、RAIDグループを対象とするデータを一時的に保持する。退避用記憶部は、例えば、RAIDグループと同一構成を有する別のRAIDグループ、1つまたは複数の論理ボリューム、1つまたは複数のディスクドライブ等として実現する

ことができる。

【0011】

第1制御部は、RAIDグループを構成する各データディスクドライブにおけるアクセスエラーの発生を監視している。アクセスエラーとしては、例えば、データの読み込みエラー、データの書き込みエラーがある。具体的なアクセスエラーとしては、例えば、ディスク面の傷のためにデータを書き込めなかった場合、ディスク面の磁性劣化でデータを読み出せなかった場合、ヘッドの故障や劣化等でデータの読み書きができなかった場合等を挙げることができる。第1制御部は、各データディスクドライブのそれぞれについて、アクセスエラーの発生を監視する。アクセスエラーの発生頻度が所定の閾値以上になった場合、第1制御部は、閾値以上のアクセスエラーが検出されたデータディスクドライブに記憶されているデータを、予備ディスクドライブにコピーさせる。ここで、注意すべき点は、アクセスエラーが閾値以上になった場合でも、実際に読み書き不能な障害が発生しているとは限らない点である。従って、第1制御部は、閾値以上のアクセスエラーが検出されたデータディスクドライブからデータを直接読み出して、予備ディスクドライブに移行させることができる。閾値以上のアクセスエラーが検出されたデータディスクドライブからデータを直接読み出せない場合、第1制御部は、他の正常なデータディスクドライブからデータ及びパリティを取り出して、データを復元し、復元したデータを予備ディスクドライブに記憶させることができる。

【0012】

第1制御部による予備ディスクドライブへのコピー処理中においても、ディスクアレイ装置を利用するホストコンピュータは、RAIDグループへアクセスし、所望のデータを読み出したり、書き込んだりすることができる。第1制御部によるコピー中に、RAIDグループを対象とする書き込み要求が発生した場合、第2制御部は、この書き込み要求を退避用記憶部に対して実行させる。即ち、新たなデータは、RAIDグループを構成する各データディスクドライブに記憶されるのではなく、退避用記憶部に記憶される。そして、第1制御部によるコピーが終了すると、第3制御部は、退避用記憶部に記憶されたデータを、閾値以上のアクセスエラーが検出されたデータディスクドライブ以外の各データディスクドライブ及び予備ディスクドライブにコピーして反映させる。

【0013】

第1制御部による予備ディスクドライブへのコピー中に、RAIDグループを構成する各データディスクドライブに対して、データの読み出し要求が発生する場合もある。第2制御部は、閾値以上のアクセスエラーが検出されたデータディスクドライブを対象とする読み出し要求が発生した場合、この閾値以上のデータディスクドライブ以外の各データディスクドライブに記憶されたデータから、要求されたデータを復元することができる。第2制御部は、復元したデータを読み出し要求元に提供する。

【0014】

逆に、閾値以上のアクセスエラーが検出されたデータディスクドライブ以外の各データディスクドライブを対象とする読み出し要求が発生した場合、第2制御部は、退避用記憶部に記憶されたデータを読み出して、この読み出したデータを読み出し要求元に提供することができる。

【発明を実施するための最良の形態】

【0015】

以下、図1～図29に基づき、本発明の実施の形態を説明する。本実施形態では、以下のようないくつかの特徴を備えることができる。

1つの態様では、予備ディスクドライブへのデータ移行中にアクセス要求を処理する第2制御部を、退避用記憶部に書き込まれたデータを管理するための差分管理情報に連付ける。第2制御部は、差分管理情報に基づいて、ホストコンピュータからの読み出し要求に対応する記憶領域を判別する。差分管理情報に記録されているデータの読み出しが要求された場合、第2制御部は、要求されたデータを退避用記憶部から読み出してホストコンピュータに提供する。逆に、差分管理情報に記録されていないデータの読み出しが要求された場

合、第2制御部は、閾値以上のデータディスクドライブ以外の各データディスクドライブに記憶されたデータに基づいてデータを復元し、この復元したデータをホストコンピュータに提供する。

【0016】

1つの態様では、第2制御部は、RAIDグループを対象とする書き込み要求のうち閾値以上のアクセスエラーが検出されたデータディスクドライブへの書き込み要求のみを退避用記憶部に対して実行させる。閾値以上のアクセスエラーが検出されたデータディスクドライブ以外の各データディスクドライブへの書き込み要求である場合、第2制御部は、当該各データディスクドライブに対して実行させる。

【0017】

1つの態様では、第2制御部は、退避用記憶部に所定値以上の空き容量がある場合に、RAIDグループを対象とする書き込み要求を退避用記憶部に対して実行させる。退避用記憶部に所定値以上の空き容量が無い場合、第2制御部は、RAIDグループを対象とする書き込み要求を、RAIDグループに対して実行させる。

【0018】

1つの態様では、第1制御部は、閾値以上のアクセスエラーが検出されたデータディスクドライブ以外の各データディスクドライブ内に記憶されたデータに基づいて、閾値以上のアクセスエラーが検出されたデータディスクドライブ内のデータを復元する。第1制御部は、復元されたデータを予備ディスクドライブにコピーさせる。

【0019】

1つの態様では、第1制御部によるコピー処理実行させる手動指示部を設けている。即ち、アクセスエラーが所定の閾値に達していない場合でも、システム管理者等は、手動指示部を介して、RAIDグループを構成するいずれかのデータディスクドライブの記憶内容を予備ディスクドライブにコピーさせることができる。

【0020】

1つの態様では、第1制御部及び第2制御部は多重動作可能となっている。そして、退避用記憶部は、複数のRAIDグループのそれぞれを対象とする書き込み要求を受け入れるようになっている。

【0021】

また、本実施形態は、例えば、ディスクアレイ装置の障害回避方法として捉えることも可能である。即ち、本実施形態は、RAIDグループを構成する複数のデータディスクドライブと、これら各データディスクドライブの予備として少なくとも1つ設けられる予備ディスクドライブと、各データディスクドライブ及び予備ディスクドライブとは別に設けられる退避用記憶部とを含んだディスクアレイ装置の障害回避方法であって、以下の第1ステップ～第5ステップを備える。第1ステップは、各データディスクドライブに対するアクセスエラーの発生を監視し、アクセスエラーの発生頻度が予め設定された所定の閾値以上になったか否かを判定する。第2ステップは、第1ステップにより閾値以上のデータディスクドライブが検出された場合、この閾値以上のデータディスクドライブに記憶されたデータを予備ディスクドライブにコピーさせる。第3ステップは、第1ステップによるコピーの開始によって、RAIDグループと退避用記憶部とを関連付ける。第4ステップは、第1ステップによるコピー中に、RAIDグループを対象とするアクセス要求が発生したか否かを判定する。第5ステップは、第4ステップによりアクセス要求の発生が検出された場合、アクセス要求が書き込み要求であるならば、第3ステップにより関連付けられた退避用記憶部に対してデータを書き込む。

【0022】

さらに、本実施形態は、例えば、ディスクアレイ装置のディスクドライブ使用方法として捉えることもできる。即ち、本実施形態は、RAIDグループを構成する複数のディスクドライブを含んだディスクアレイ装置のディスクドライブ使用方法であって、以下のステップを備える。障害ドライブ検出ステップは、RAIDグループを構成する各ディスクドライブに対するアクセスエラーの発生を監視し、アクセスエラーの発生頻度が予め設定

された所定の閾値以上になった場合に障害ディスクドライブであると判定する。データコピーステップは、障害ドライブ検出ステップによって障害ディスクドライブが検出された場合は、この障害ディスクドライブに記憶されたデータを、RAIDグループを構成する各ディスクドライブ以外の正常ディスクドライブにコピーさせる。アクセス要求検出ステップは、データコピーステップによるコピー中に、RAIDグループを対象とするアクセス要求が発生したか否かを検出する。アクセス処理ステップは、アクセス要求検出ステップにより書き込み要求が検出された場合は、データコピーがされている正常ディスクドライブとは別の正常ディスクドライブに対して、書き込み要求に係わるデータを書き込む。

【実施例1】

【0023】

図1～図9に基づいて、本発明の第1実施例を説明する。図1は、ディスクアレイ装置10の概略構成を示すブロック図である。

ディスクアレイ装置10は、通信ネットワークCN1を介して、複数のホストコンピュータ1と双方向通信可能に接続されている。ここで、通信ネットワークCN1は、例えば、LAN (Local Area Network)、SAN (Storage Area Network)、インターネット等である。LANを用いる場合、ホストコンピュータ1とディスクアレイ装置10との間のデータ転送は、TCP/IP (Transmission Control Protocol/Internet Protocol) プロトコルに従って行われる。SANを用いる場合、ホストコンピュータ1とディスクアレイ装置10とは、ファイバチャネルプロトコルに従ってデータ転送を行う。また、ホストコンピュータ1がメインフレームの場合は、例えば、FICON (Fibre

Connection : 登録商標)、ESCON (Enterprise

System Connection : 登録商標)、ACONARC (Advanced

Connection Architecture : 登録商標)、FIBARC (Fibre

Connection Architecture : 登録商標) 等の通信プロトコルに従ってデータ転送が行われる。

【0024】

各ホストコンピュータ1は、例えば、サーバ、パーソナルコンピュータ、ワークステーション、メインフレーム等として実現されるものである。例えば、各ホストコンピュータ1は、図外に位置する複数のクライアント端末と別の通信ネットワークを介して接続されている。各ホストコンピュータ1は、例えば、各クライアント端末からの要求に応じて、ディスクアレイ装置10にデータの読み書きを行うことにより、各クライアント端末へのサービスを提供する。

【0025】

ディスクアレイ装置10は、それぞれ後述するように、各チャネルアダプタ（以下、CHAと略記）11と、各ディスクアダプタ（以下、DKAと略記）12と、共有メモリ13と、キャッシュメモリ14と、スイッチ部15と、各ディスクドライブ16とを備えて構成されている。CHA11及びDKA12は、例えば、プロセッサやメモリ等が実装されたプリント基板と、制御プログラムとの協働により実現される。

【0026】

ディスクアレイ装置10には、例えば、4個や8個等のように、複数のCHA11が設けられている。チャネルアダプタ11は、例えば、オープン系用CHA、メインフレーム系用CHA等のように、ホストコンピュータ1の種類に応じて、用意される。各CHA11は、ホストコンピュータ1との間のデータ転送を制御するものである。各CHA11は、それぞれプロセッサ部、データ通信部及びローカルメモリ部を備えている（いずれも不図示）。

【0027】

各CHA11は、それぞれに接続されたホストコンピュータ1から、データの読み書きを要求するコマンド及びデータを受信し、ホストコンピュータ1から受信したコマンドに従って動作する。DKA12の動作も含めて先に説明すると、例えば、CHA11は、ホ

ストコンピュータ1からデータの読み出し要求を受信すると、読み出しコマンドを共有メモリ13に記憶させる。DKA12は、共有メモリ13を随時参照しており、未処理の読み出しコマンドを発見すると、ディスクドライブ16からデータを読み出して、キャッシングメモリ14に記憶させる。CHA11は、キャッシングメモリ14に移されたデータを読み出し、コマンド発行元のホストコンピュータ1に送信する。また例えば、CHA11は、ホストコンピュータ1からデータの書き込み要求を受信すると、書き込みコマンドを共有メモリ13に記憶させると共に、受信データをキャッシングメモリ14に記憶させる。DKA12は、共有メモリ13に記憶されたコマンドに従って、キャッシングメモリ14に記憶されたデータを所定のディスクドライブ16に記憶させる。

【0028】

各DKA12は、ディスクアレイ装置10内に例えば4個や8個等のように複数個設けられている。各DKA12は、各ディスクドライブ16との間のデータ通信を制御するもので、それぞれプロセッサ部と、データ通信部と、ローカルメモリ等を備えている（いずれも不図示）。各DKA12と各ディスクドライブ16とは、例えば、SAN等の通信ネットワークCN2を介して接続されており、ファイバチャネルプロトコルに従ってブロック単位のデータ転送を行う。各DKA12は、ディスクドライブ16の状態を随時監視しており、この監視結果は内部ネットワークCN3を介してSVP2に送信される。

【0029】

ディスクアレイ装置10は、多数のディスクドライブ16を備えている。ディスクドライブ16は、例えば、ハードディスクドライブ(HDD)や半導体メモリ装置等として実現される。ここで、例えば、4個のディスクドライブ16によってRAIDグループ17を構成することができる。RAIDグループ17とは、例えばRAID5(RAID5に限定されない)に従って、データの冗長記憶を実現するディスクグループである。各RAIDグループ17により提供される物理的な記憶領域の上には、論理的な記憶領域である論理ボリューム18(LU)を少なくとも1つ以上設定可能である。

【0030】

「制御メモリ」の一例に該当する共有メモリ13は、例えば、不揮発メモリによって構成されており、制御情報や管理情報を記憶する。キャッシングメモリ14は、主としてデータを記憶する。

【0031】

SVP(Service Processor)2は、ディスクアレイ装置10の管理及び監視を行うためのコンピュータ装置である。SVP2は、ディスクアレイ装置10内に設けられた通信ネットワークCN3を介して、各CHA11及び各DKA12等から各種の環境情報や性能情報を収集する。SVP2が収集する情報としては、例えば、装置構成、電源アラーム、温度アラーム、入出力速度(IOPS)等が挙げられる。通信ネットワークCN3は、例えば、LANとして構成される。システム管理者は、SVP2の提供するユーザインターフェースを介して、RAID構成の設定、各種パッケージ(CHA、DKA、ディスクドライブ等)の閉塞処理等を行うことができる。

【0032】

図2は、ディスクアレイ装置10内に記憶されるRAID構成管理テーブルT1の概略構造を示す説明図である。RAID構成管理テーブルT1は、例えば共有メモリ13内に記憶される。RAID構成管理テーブルT1は、例えば、RAIDグループ番号(図中、グループ#)と、論理ボリューム番号(図中、ボリューム#)と、ディスクドライブ番号(図中、ディスク#)と、RAIDレベルとを対応付けている。以下に述べる他のテーブルも同様であるが、テーブル内の文字または数値は、説明のためのものであって、実際に記憶されるものとは異なる。RAID構成管理テーブルT1の内容の一例を説明すると、例えば、グループ番号1のRAIDグループ17には、ボリューム番号1～3の合計3個の論理ボリューム18が設定されている。また、このRAIDグループ17は、ディスク番号1～4で特定される合計4個のディスクドライブ16から構成されている。そして、このグループ番号1で特定されるRAIDグループ17は、RAID5に従って運用され

ている。

【0033】

本実施例では、後述のように、あるディスクドライブ16に障害発生の予兆が検出された場合、この障害発生が予測されるディスクドライブ16が所属するRAIDグループへのデータ書き込みを、他のRAIDグループ（あるいは、論理ボリュームやディスクドライブ）に退避させるようになっている。

【0034】

図2(a)は、退避用のRAIDグループ17を設定する前の構成を示し、図2(b)は、退避用のRAIDグループ17を設定した後の構成を示す。図2(a)に示すように、グループ番号5で特定されるRAIDグループ17は、当初使用目的が設定されておらず、論理ボリュームが1つも設定されていない。グループ番号1のRAIDグループ17に属するいずれか1つのディスクドライブ16に障害の発生が予測されると、グループ番号5で特定される未使用のRAIDグループ17は、退避用のRAIDグループ17として利用される。データ退避用に使用されるRAIDグループ17(#5)には、データ退避元のRAIDグループ17(#1)に設定されている論理ボリューム18(#1~3)と同数の論理ボリューム(#13~15)が設定される。

【0035】

図3は、ディスクアレイ装置10内に記憶されるペア情報管理テーブルT2の概略構造を示す説明図である。ペア情報管理テーブルT2は、例えば、共有メモリ13内に記憶されるもので、ペアを構成する論理ボリューム18について管理する。

【0036】

ペア情報管理テーブルT2は、例えば、正ボリューム番号と、副ボリューム番号と、ペア状態と、差分ビットマップとを対応付けている。図3(a)に示すペア情報管理テーブルT2は、データ退避用の論理ボリューム18を設定する前の状態を示している。図3(a)では、例えば、ある1つの論理ボリューム18(#4)が正、別の1つの論理ボリューム18(#7)が副となってペアを構成している。ペア状態は「二重化」である。二重化とは、正ボリュームと副ボリュームとの記憶内容を同期させることを意味する。差分ビットマップについてはさらに後述するが、正ボリュームと副ボリュームとのデータの差分を管理するための情報である。

【0037】

図3(b)は、データ退避用のRAIDグループ17を設定した場合を示す。RAIDグループ17(#1)の各論理ボリューム18(#1~3)は、RAIDグループ17(#5)に設定された各論理ボリューム18(#13~15)にそれぞれ一対一で対応付けられる。即ち、図3(b)に示す例では、論理ボリューム18(#1)は、論理ボリューム18(#13)とペアを構成し、論理ボリューム18(#2)は、論理ボリューム18(#14)とペアを構成し、論理ボリューム18(#3)は、論理ボリューム18(#15)とペアを構成する。これらの各ペアのペア状態は、「二重化」ではなく、「更新データ退避中」となっている。「更新データ退避中」とは、データ退避元の論理ボリューム18(#1~3)を対象とする更新データを、データ退避先の論理ボリューム18(#13~15)に退避させている状態を示す。「更新データ退避中」状態と、「二重化」状態とでは、例えば、初期コピーを行わない点で相違する。通常の二重化では、最初に初期コピーを行って、正ボリュームと副ボリュームとの内容を一致させるが、「更新データ退避中」状態では、初期コピーを行わない。

【0038】

図4は、差分ビットマップ20について説明する説明図である。図4(a)に示すように、本実施形態では、正ボリュームと副ボリュームとでペアを形成し、正ボリュームへデータ書き込み(更新)が要求された場合は、このデータを副ボリュームに記憶させるようになっている。仮に、データ(#1)とデータ(#2)の更新があった場合、これらのデータは、副ボリュームに記憶される。そして、更新データに対応する差分ビットには、それぞれ「1」がセットされる。差分ビットに「1」がセットされた状態は、副ボリューム内

のデータが正ボリュームに反映されていないこと、即ち、新たなデータが副ボリュームに記憶されていることを意味する。従って、データ読出し要求があった場合、要求されたデータに対応する差分ビットが「1」にセットされているならば、そのデータは、副ボリュームに記憶されていると判別することができる。逆に、読出し対象のデータに対応する差分ビットが「0」にセットされているならば、要求されたデータは、正ボリュームに記憶されていると判別することができる。

【0039】

図4（b）に示すように、差分ビットマップ20は、差分ビットの集合体である。差分ビットマップ20は、「差分管理情報」の一例である。本実施例において、各差分ビットは、ディスクの各トラックにそれぞれ対応している。従って、更新管理単位は、トラック単位である。更新管理単位に満たないデータの更新がされた場合は、この更新データが属するトラックの全データをキャッシュメモリ14に読み出し、キャッシュメモリ14上で更新データと合成させる。そして、このキャッシュメモリ14上で合成されたトラックを副ボリュームに記憶させ、対応する差分ビットを「1」にセットする。

【0040】

次に、図5は、本実施例による障害回避方法の全体概要を示す説明図である。図5に示す例では、RAIDグループ17（P）に属する4番目のディスクドライブ16（#4）に障害発生が予測されたものとする。詳細は後述するが、読み出しエラーや書き込みエラーが所定の閾値以上に発生した場合、このディスクドライブ16（#4）は、障害発生のおそれありと判定される。そこで、まず最初に、障害発生が予測されたディスクドライブ16（#4）の記憶内容がキャッシュメモリ14に読み出され、キャッシュメモリ14から予備ディスクドライブ16（S P）にコピーされる（S1）。

【0041】

予備ディスクドライブ16（S P）へのデータコピーが開始されると、ディスクアレイ装置10が有する複数のRAIDグループ17のうち、未使用のRAIDグループが1つ確保される（S2）。そして、障害発生が予測されたディスクドライブ16（#4）の属するRAIDグループ17（P）を正、S2で確保された未使用のRAIDグループ17（S）を副として、ペアが形成される。正のRAIDグループ17（P）に設定されている正ボリューム18（P）と、副のRAIDグループ17（S）に設定される副ボリューム18（S）とは、ペアを形成する（S3）。このペアに関する情報は、ペア情報管理テーブルT2に登録される。

【0042】

予備ディスクドライブ16（S P）へのデータ移行中に、ホストコンピュータ1からデータ書き込みが要求された場合、このデータは、正ボリューム18（P）ではなく、副ボリューム18（S）に記憶される（S4）。副ボリューム18（S）にデータが記憶された場合、この更新データに対応する差分ビットが「1」にセットされ、差分ビットマップ20により管理される（S5）。

【0043】

予備ディスクドライブ16（S P）へのデータ移行中に、ホストコンピュータ1からデータ読出しが要求された場合、DKA12は、差分ビットマップ20を参照することにより、ホストコンピュータ1から要求されたデータが正ボリューム18（P）と副ボリューム18（S）のいずれに記憶されているかを判別する。要求されたデータに対応する差分ビットが「0」にセットされている場合、この要求されたデータは、正ボリューム18（P）に記憶されている。そこで、DKA12は、要求されたデータを正ボリューム18（P）から読み出し、キャッシュメモリ14にコピーする。CHA11は、キャッシュメモリ14に移されたデータを、ホストコンピュータ1に送信する（S6）。一方、ホストコンピュータ1から要求されたデータに対応する差分ビットが「1」にセットされている場合、この要求されたデータは、副ボリューム18（S）に存在する。そこで、DKA12は、要求されたデータを副ボリューム18（S）から読み出してキャッシュメモリ14にコピーする。前記同様に、CHA11は、キャッシュメモリ14に移されたデータをホス

トコンピュータ1に送信する（S7）。

【0044】

予備ディスク16（SP）へのデータ移行が完了すると、DKA12は、差分ビットマップ20を参照し、副ボリューム18（S）に退避したデータを、正ボリューム18（P）側に反映させる（S8）。より詳しくは、副ボリューム18（S）に記憶されたデータは、正のRAIDグループ17（P）に属するディスクドライブ16のうち、障害が予測されたディスクドライブ16（#4）以外のディスクドライブ16（#1～3）と、予備ディスクドライブ16（SP）とにコピーされる。言うまでもないが、副ボリューム18（S）に記憶されたデータの全部をディスクドライブ16（#1～3）及び予備ディスクドライブ（SP）にそれぞれコピーするのではない。対応するディスクにのみ必要なデータがコピーされる。

【0045】

次に、図5中のS1で示した予備ディスクドライブ16（SP）へのコピー処理について、図6を参照しつつ説明する。本実施例においては、予備ディスクドライブ16（SP）へのデータコピーを「スペアリング」と称する場合がある。図6に示すフローチャートは、「第1制御部」、「第1ステップ」及び「第2ステップ」、「障害ドライブ検出ステップ」及び「データコピーステップ」の一例である。図6に示す処理は、例えば、DKA12によって実行される。なお、以下の各フローチャートでも同様であるが、各フローチャートは処理の概要を示すもので、実際のコンピュータプログラムとは相違する。

【0046】

DKA12は、各ディスクドライブ16におけるアクセスエラー（IOエラー）を監視している（S11）。エラー発生が検出された場合（S11:YES）、DKA12は、エラー種別毎にエラー発生回数を管理する（S12）。DKA12は、例えば、図6中に示すエラー管理テーブルT3を用いることにより、発生したアクセスエラーを管理することができる。アクセスエラーは、その種類（ET1～ET3・・・）毎に発生回数（N1～N3・・・）が管理され、かつ、アクセスエラーの種類毎に閾値Th1～Th3・・・がそれぞれ設定されている。図6中では1つだけ図示するが、エラー管理は、使用されている各ディスクドライブ16毎にそれぞれ行われる。

【0047】

ここで、アクセスエラーは、例えば、読み出しエラーと書き込みエラーとに分類することができる。また、アクセスエラーは、例えば、リカバリ可能なエラーとリカバリ不可能なエラーとに分類することもできる。リカバリ可能なエラーとは、例えば、ECC（Error-Correcting Code）によりデータの修復を容易に行える種類のエラーを意味する。リカバリ不可能なエラーとは、各データに付加された冗長データ（ECC）ではエラーを修復することができず、より上位での回復（他のデータとパリティによる逆演算等）が必要となる種類のエラーを意味する。アクセスエラーの具体例としては、例えば、ディスク面に物理的な傷が存在するためにデータを書き込むことができない場合、ディスク面の磁性が劣化しているためデータを読み出すことができない場合、磁気ヘッドの不良でデータの読み書きができない場合等を挙げることができる。

【0048】

エラー管理テーブルT3の下側に示すように、リカバリ可能なエラーとリカバリ不可能なエラーとでは、閾値Thの設定値が異なる。リカバリ可能なエラーの閾値Thは、相対的に高く設定され、リカバリ不可能なエラーの閾値Thは、相対的に低く設定される。なお、図6中のエラー管理テーブルT3では、3種類以上のエラーを示し、各種類のエラー毎にそれぞれ閾値Thを設定しているが、これは一例であって、リカバリ可能エラー及びリカバリ不可能エラーの2種類に限定してもよい。あるいは、さらに詳しくエラーを分類し、エラー管理テーブルT3に示すように、多種類のエラー毎にそれぞれ閾値Thを設定するようにもよい。

【0049】

DKA12は、エラー管理テーブルT3を参照することにより、使用されているディス

クドライブ16のそれぞれについて、アクセスエラーの発生頻度が閾値T h以上になったか否かを判定する（S13）。アクセスエラーの発生頻度が閾値T h以上になつてない場合は（S13: NO）、処理を終了する。一方、アクセスエラーの発生頻度が閾値T h以上になつた場合は（S13: YES）、そのディスクドライブ16に障害の発生が予測された場合である。そこで、DKA12は、障害の発生が予測されたディスクドライブ（以下、このドライブを障害ディスクドライブと称する場合がある）16の記憶内容を、予備ディスクドライブ16（SP）にコピーし、データ移行を開始させる（S14）。データ移行が完了するまで（S15: NO）、S14の処理が繰り返される。予備ディスクドライブ16（SP）へのデータ移行が完了すると（S15: YES）、処理を終了する。

【0050】

なお、上記処理では、エラー種別毎にそれぞれ閾値T hを設定し、いずれかの種類のアクセスエラーの発生頻度が、それに対応する閾値T h以上になつた場合に、障害ディスクドライブであると判定している。しかし、これに限らず、アクセスエラーを総合的に解析することにより（アクセスエラーに基づいて）、障害ディスクドライブであるか否かを判定してもよい。

【0051】

図7は、SVP2を介して、手動操作によりスペアリングを実行させる場合の処理を示す。図7に示す処理は、主としてSVP2とDKA12との協働作業により実行される。この処理は、「手動指示部」に対応する構成を含んでいる。

【0052】

SVP2は、内部ネットワークCN3を介して、各DKA12から各ディスクドライブ16に関するエラー情報を収集している（S21）。SVP2は、システム管理者からの要求に応じて、あるいは自動的に、収集したエラー情報をSVP2の端末画面に表示させる（S22）。SVP2は（より正確には、SVP2のマイクロプロセッサにより実行される制御プログラムは）、各ディスクドライブ16のそれぞれについて、アクセスエラーの発生頻度が閾値T h以上になつたか否かを判定する（S23）。アクセスエラーの発生頻度が閾値T h以上になつたディスクドライブ16が検出された場合（S23: YES）、SVP2は、このディスクドライブ16を将来障害の発生する可能性が高い障害ディスクドライブであると判定し、システム管理者に警告する（S24）。この警告は、例えば、警告メッセージの表示または音声出力、警告ランプの点滅等により行うことができる。アクセスエラーの発生頻度が閾値T h以上になつたディスクドライブ16が存在しない場合（S23: NO）、S24はスキップされる。

【0053】

システム管理者は、S24で通知された警告に従って、あるいは、警告がされていない場合でも自らの判断に従って、スペアリングの開始を指示できる。システム管理者からの手動操作によるスペアリング開始指示は、SVP2のユーザインターフェース（例えば、キーボードスイッチからの入力や音声による指示等）により行われる。DKA12は、システム管理者からのスペアリングの開始指示があったか否かを判定する（S25）。手動操作による開始指示が無い場合（S25: NO）、処理を終了するか否かを判定する（S26）。例えば、システム管理者がメニュー操作等を行うことにより処理の終了を指示した場合（S26: YES）、処理は終了する。システム管理者が処理の終了を指示しない場合（S26: NO）、S21に戻ってエラー情報の収集等が繰り返される。

【0054】

システム管理者の手動操作によってスペアリングの開始が指示された場合（S25: YES）、システム管理者により指示されたディスクドライブ16またはS24で警告されたディスクドライブ16、あるいはシステム管理者により指示されたディスクドライブ16及び警告されたディスクドライブ16の記憶内容が、予備ディスクドライブ16（SP）にコピーされる（S27）。そして、予備ディスクドライブ16（SP）へのデータ移行が完了すると（S28: YES）、処理を終了する。

【0055】

図8は、データ退避処理を示すフローチャートである。データ退避処理は、スペアリングの開始により起動されるもので、DKA12によって実行される。図8に示す処理は、「第2制御部」、「第3ステップ」～「第5ステップ」、「アクセス要求検出ステップ」及び「アクセス処理ステップ」にそれぞれ対応する一例である。

【0056】

DKA12は、スペアリング、即ち、障害ディスクドライブ16から予備ディスクドライブ16(SP)へのデータコピーが開始されたか否かを監視している(S31)。スペアリング開始が検出されると(S31:YES)、DKA12は、未使用のRAIDグループ17が存在するか否かを判定する(S32)。未使用のRAIDグループ17が存在しない場合(S32:NO)、データ退避領域を確保できないので、処理を終了する。

【0057】

未使用のRAIDグループ17を発見した場合(S32:YES)、DKA12は、障害ディスクドライブ16が属するRAIDグループ17を正、発見された未使用のRAIDグループ17を副として、ペアを構成する(S33)。正のRAIDグループ17に複数の論理ボリューム18が設定されている場合、副のRAIDグループ17にも同数かつ同サイズの論理ボリューム18がそれぞれ設定され、正と副の各論理ボリューム18同士でペアが形成される。

【0058】

DKA12は、隨時共有メモリ13を参照することにより、ホストコンピュータ1からのアクセス要求(読み出し要求または書き込み要求)が発生したか否かを監視している(S34)。ホストコンピュータ1からのアクセス要求が発生していない場合(S34:NO)、DKA12は、スペアリングが終了したか否かを判定する(S35)。スペアリングが終了していない場合(S35:NO)、S34に戻る。スペアリングが終了した場合(S35:YES)、DKA12は、副ボリューム18に記憶されたデータを、正ボリューム18に反映させ(S36)、ボリュームペアを解除し(S37)、処理を終了する。

【0059】

スペアリング中にホストコンピュータ1からのアクセス要求が発生した場合(S34:YES)、DKA12は、このアクセス要求が読み出し要求(図中、リードと表示)であるか否かを判定する(S38)。読み出し要求である場合(S38:YES)、DKA12は、差分ビットマップ20を参照し、読み出しを要求されたデータに対応する差分ビットに「1」がセットされているか否か(図中では、差分ビットに1をセットする場合をON、差分ビットに0をセットする場合をOFFと示す)を判定する(S39)。

【0060】

差分ビットに「1」がセットされている場合(S39:YES)、要求されたデータは副ボリューム18に存在する。そこで、DKA12は、副ボリューム18からデータを読み出して、キャッシュメモリ14に格納する(S40)。読み出しを要求されたデータに対応する差分ビットに「0」がセットされている場合(S39:NO)、要求されたデータは正ボリューム18に存在するので、DKA12は、正ボリューム18からデータを読み出し、キャッシュメモリ14に格納する(S41)。ここで、要求されたデータが障害ディスクドライブ16に記憶されている場合は、障害ディスクドライブ16から直接データを読み出すのではなく、他の正常なディスクドライブ16に格納されているデータに基づいて、要求されたデータを復元する。

【0061】

ホストコンピュータ1からのアクセス要求が書き込み要求である場合(S38:NO)、DKA12は、書き込みデータ(更新データ)に対応する差分ビットに「1」をセットし(S42)、書き込みデータを副ボリューム18に記憶させる(S43)。

【0062】

図9は、差分データのフィードバック処理を示すフローチャートである。差分データフィードバック処理は、スペアリングの終了により、DKA12によって実行される。本処理は、図8中のS36の詳細である。本処理は、「第3制御部」、「第6ステップ」、「

データ更新ステップ」に対応する一例である。

【0063】

DKA12は、フィードバックポインタを論理ボリュームの先頭アドレスにセットする(S51)。DKA12は、そのアドレスに対応する差分ビットに「1」がセットされているか否かを判定する(S52)。差分ビットに「1」がセットされている場合(S52:Y ES)、DKA12は、そのアドレスのデータを副ボリューム18から正ボリューム18にコピーさせる(S53)。より詳しくは、副ボリューム18から読み出されたデータは、キャッシュメモリ14にコピーされ、キャッシュメモリ14から正ボリューム18にコピーされる。1アドレス分のデータコピーを終了すると、DKA12は、フィードバックポインタを次のアドレスに移動させる(S54)。そして、DKA12は、差分データのフィードバックが完了したか否かを判定する(S55)。即ち、DKA12は、フィードバックポインタが最終位置を示しているか否かを判定する。差分データのフィードバックが完了するまで(S55:N0)、S52～S54の処理が繰り返し実行される。

【0064】

このように構成される本実施例によれば、以下の効果を奏する。

障害ディスクドライブ(正確には、障害の発生が予測されるディスクドライブ)16から予備ディスクドライブ16(SP)へのデータ移行中に、障害ディスクドライブ16の属するRAIDグループ17へのデータ読み書きを低減することができる。従って、RAIDグループ17を構成する他の正常なディスクドライブ16に障害が発生する可能性を少なくすることができ、いわゆる二重障害の可能性を低減できる。即ち、例えば、RAID5に従う一組のデータセットを考えた場合、このデータセットを構成するいずれか1つのデータが失われた場合でも、残りのデータ(パリティを含む)から逆演算を行うことにより、消失したデータを復元することができる。具体的には、例えば、データセットが、D1～D4の4個のデータと1個のパリティpとから構成される場合を考える。仮に、D2の読み出しを行ふことができない場合、D2は、 $D_2 = (D_1) \text{XOR} (D_3) \text{XOR} (D_4) \text{XOR} (p)$ よって求めることができる。しかし、2つ以上のデータが利用できない場合、演算によるデータ復元は不可能である。

【0065】

障害ディスクドライブ16であると判定されていない他の正常なディスクドライブ16であっても、全くエラーが存在しないとは限らない。また、エラーが全く存在しない場合でも、アクセス回数が多くなるほどエラーを生じる確率が増す。もしも、正常なディスクドライブ16に発生したエラーの位置が、障害ディスクドライブ16のエラー位置と偶然一致した場合、その場所に格納されているデータを復元することはできない。障害ディスクドライブ16には比較的多数のエラーが既に生じているので、正常なディスクドライブ16に追加的に発生した新たなエラーの位置が、障害ディスクドライブ16のエラー位置と偶然一致するおそれがある。このようなエラー位置の一致による障害を本実施例では「二重障害」と呼ぶ。従って、スペアリングの最中に、正常なディスクドライブ16へのアクセスを通常通り続行すると、正常なディスクドライブ16に生じた新たなエラーによって、データの一部を消失する可能性がある。

【0066】

これに対し、本実施例では、スペアリング中に、正常な他のディスクドライブ16へのアクセスを低減するため、正常な他のディスクドライブ16に新たなエラーが追加的に発生して、二重障害が発生する可能性を少なくできる。具体的には、本実施例では、スペアリング中におけるデータ書き込みは、副ボリューム18に対して行わせ、スペアリング中のデータの読み出しは、要求されたデータが正ボリューム18に存在する場合に限って、正ボリューム18から読み出す。従って、障害ディスクドライブ16が属する正ボリューム18へのアクセス頻度を低減し、二重障害の発生を防止可能である。

【0067】

また、本実施例では、副ボリューム18に退避させたデータを差分ビットマップ20によって管理する。従って、ホストコンピュータ1からデータの読み出し要求があった場合に

、要求されたデータが正ボリューム18または副ボリューム18のいずれに存在するかを容易に判別することができる。

【実施例2】

【0068】

図10～図13に基づいて、本発明の第2実施例を説明する。本実施例の1つの特徴は、スペアリング中のデータ退避領域として、論理ボリュームを使用する点にある。また、本実施例の1つの特徴は、ジャーナルファイルを使用する点にある。図10は、本実施例による障害回避方法の全体動作の概要を示す説明図である。動作全体の概要は、前記実施例とほぼ同様である。

【0069】

あるディスクドライブ16について障害の発生が予測されると、この障害発生が予測されたディスクドライブ16の記憶内容を予備ディスクドライブ16(SP)に移行させるスペアリングが開始される(S61)。スペアリングが開始されると、データ退避領域用に、未使用の論理ボリューム18が少なくとも1つ以上確保される(S62)。この未使用の論理ボリューム18は、ワークボリューム18(W)として利用される。ここで、注意すべき点は、前記実施例とは異なり、同サイズの未使用RAIDグループを確保するのではなく、未使用の論理ボリュームを確保する点である。即ち、データ退避元の記憶サイズとデータ退避先の記憶サイズとが相違し、データ退避元よりも小さな記憶サイズを有するデータ退避先を使用する点である。

【0070】

データ退避元のRAIDグループ17(P)に設定された論理ボリューム18(P)と、ワークボリューム18(W)とが対応付けられる(S63)。論理ボリューム18(P)とワークボリューム18(W)とは、記憶サイズが異なってもよい(同一サイズであってもよい)。ホストコンピュータ1からRAIDグループ17(P)に対する書き込み要求が発生すると、この更新データは、ワークボリューム18(W)に順次記憶されていく(S64)。ここで注意すべき点は、ワークボリューム18(W)には、ジャーナルファイルのように、書き込みの履歴が記憶される点である。

【0071】

ホストコンピュータ1からRAIDグループ17(P)に対する読み出し要求が発生した場合、要求されたデータがRAIDグループ17(P)に存在するならば、つまり、更新されていないデータの読み出し要求の場合は、論理ボリューム18(P)からデータが読み出され、キャッシュメモリ14及びCHA11等を介して、ホストコンピュータ1に提供される(S65)。要求されたデータが障害ディスクドライブ16(#4)に存在する場合、他のディスクドライブ16(#1～3)が記憶するデータに基づいて、要求されたデータが復元される。

【0072】

ホストコンピュータ1から要求されたデータがワークボリューム18(W)に存在するならば、つまり、更新されたデータの読み出し要求の場合は、ワークボリューム18(W)からデータが読み出され、ホストコンピュータ1に提供される(S66)。そして、スペアリングが終了すると、ワークボリューム18(W)に記憶されたデータが、論理ボリューム18(P)及び予備ディスクドライブ16(SP)に反映される(S67)。障害ディスクドライブ16(#4)に差分データは反映されない。

【0073】

図11は、ディスクアレイ装置10内に記憶されるワークボリューム管理テーブルT4の概略構造を示す説明図である。ワークボリューム管理テーブルT4は、例えば、共有メモリ13内に記憶される。なお、前記実施例で述べた各テーブルも含めて、全てのテーブルは、共有メモリ以外の記憶領域に記憶させることもできる。

【0074】

ワークボリューム管理テーブルT4は、例えば、ワークボリューム番号と、ワークボリュームの記憶容量と、対応付けられている正ボリュームの番号と、最新のデータ更新を記

憶する終端アドレスと、差分ビットマップとを対応付けて構成されている。差分ビットマップは、更新されたデータの位置を管理するために用いられる。図11(a)は、予備ディスクドライブ16(S P)へのデータ移行(スペアリング処理)が開始される前の状態を示す。従って、3個のワークボリューム18(#10~12)は、いずれも正ボリュームに対応付けられていない。

【0075】

図11(b)は、スペアリング処理が開始された後の状態を示す。この例では、各ワークボリューム18(#10~12)を、それぞれ1つずつの正ボリューム18(#1~3)に対応付けている。しかし、これに限らず、1つのワークボリューム18を複数の正ボリューム18に対応付ける構成でもよい。

【0076】

図11(c)は、ワークボリューム18に記憶されるデータの概略構造を示す。ワークボリューム18内では、例えば、ジャーナルアドレスと、正ボリューム番号と、アドレスと、更新データとが対応付けられて記憶されている。図示の例では、上から順番にデータが記憶されていくようになっており、最下端が終端アドレスとなっている。

【0077】

図12は、データ退避処理を示すフローチャートである。本処理は、DKA12によって実行される。DKA12は、障害ディスクドライブ16から予備ディスクドライブ16(S P)へのデータコピーが開始されたことを検出すると(S71)、ワークボリューム18が登録されているか否かを判定する(S72)。ワークボリューム18が登録されていない場合(S72: NO)、データ退避領域を確保できないので処理を終了する。

【0078】

ワークボリューム18が登録されている場合(S72: YES)、登録されているワークボリューム18が未使用であるか否かを判定する(S73)。そのワークボリューム18が使用中の場合(S73: NO)、他にワークボリューム18が登録されているか否かを判定する(S74)。登録済のワークボリュームが存在しない場合(S74: NO)、データ退避領域を確保できないので処理を終了する。一方、他のワークボリューム18が登録されている場合(S74: YES)、S73に戻って未使用のワークボリューム18であるか否かを検査する。

【0079】

このようにして、DKA12は、登録されているワークボリューム18を順番に検査し、未使用のワークボリューム18を検出する。未使用のワークボリューム18が検出された場合(S73: YES)、DKA12は、この検出された未使用のワークボリューム18とデータ退避元の論理ボリューム18とを対応付けて、ワークボリューム管理テーブルT4に登録する(S75)。

【0080】

DKA12は、予備ディスクドライブ16(S P)へのデータ移行が完了するまでの期間(S77)、ホストコンピュータ1からのアクセス要求が発生したか否かを監視する(S76)。データ移行が完了した場合(S78: YES)、DKA12は、ワークボリューム18に記憶されたデータを正ボリューム18及び予備ディスクドライブ16(S P)に反映させる(S78)。そして、DKA12は、ワークボリューム管理テーブルT4から、ワークボリューム18に対応付けた正ボリューム18の番号を削除し、データ退避領域として使用したワークボリューム18を解放する(S79)。

【0081】

データ移行期間内にホストコンピュータ1からのアクセス要求が検出された場合(S76: YES)、DKA12は、このアクセス要求が読み出し要求であるか否かを判定する(S80)。読み出し要求の場合(S80: YES)、ワークボリューム管理テーブルT4に登録されている差分ビットマップを参照し、要求されたデータに対応する差分ビットに「1」が設定されているか否かを判定する(S81)。差分ビットに「1」がセットされている場合(S81: YES)、読み出すべきデータはワークボリューム18に記憶されている。そこで、D

K A 1 2 は、ワークボリューム 1 8 に記憶されたジャーナルファイルを、終端アドレスから上に向けて（古い方に遡って）順番に検索することにより、目的のデータを発見する（S 8 2）。D K A 1 2 は、発見したデータをワークボリューム 1 8 から読み出して、キャッシュメモリ 1 4 に記憶させ（S 8 3）、S 7 7 に戻る。ホストコンピュータ 1 から読み出しを要求されたデータに対応する差分ビットに「0」がセットされている場合（S81：NO）、D K A 1 2 は、目的のデータを正ボリューム 1 8 から読み出してキャッシュメモリ 1 4 に記憶させる（S 8 4）。C H A 1 1 は、キャッシュメモリ 1 4 に記憶されたデータを読み出し、ホストコンピュータ 1 に送信する。

【0082】

一方、ホストコンピュータ 1 からのアクセス要求が書き込み要求の場合（S80：NO）、D K A 1 2 は、ワークボリューム 1 8 の残量検査を行う（S 8 5）。更新データを記憶するだけの残容量がワークボリューム 1 8 に存在しない場合（S85：NO）、D K A 1 2 は、更新データを正ボリューム 1 8 に記憶させる（S 8 6）。そして、更新データを正ボリューム 1 8 に記憶させたため、更新データに対応する差分ビットに「0」を設定し（S 8 7）、S 7 7 に戻る。更新データを記憶するだけの残容量がワークボリューム 1 8 に存在する場合（S85：YES）、D K A 1 2 は、更新データに対応する差分ビットに「1」をセットし（S 8 8）、更新データをワークボリューム 1 8 に記憶させる（S 8 9）。そして、D K A 1 2 は、ワークボリューム管理テーブル T 4 の終端アドレスを更新し（S 9 0）、S 7 7 に戻る。

【0083】

なお、ワークボリューム 1 8 の残量が不足している場合（S85：NO）、S 7 2 ～ S 7 4 で行ったように、他の空いているワークボリューム 1 8 を探索し、他の空いているワークボリューム 1 8 を発見した場合は、このワークボリューム 1 8 に更新データを記憶させるようにしてもよい。

【0084】

図 1 3 は、差分データのフィードバック処理を示すフローチャートである。この処理は、図 1 2 中の S 7 8 に対応する。

【0085】

D K A 1 2 は、ワークボリューム 1 8 に退避しているデータが存在するか否かを判定する（S 1 0 0）。退避データが存在しない場合（S100：NO）、処理を終了する。退避データが存在する場合（S100：YES）、フィードバックポインタをワークボリューム 1 8 の終端アドレスにセットする（S 1 0 1）。即ち、最新のデータにフィードバックポインタをセットする。次に、D K A 1 2 は、フィードバックポインタの示すジャーナルファイル（更新データ及びアドレス）をキャッシュメモリ 1 4 に記憶させる（S 1 0 2）。D K A 1 2 は、キャッシュメモリ 1 4 にコピーされた更新データを正ボリューム 1 8 にコピーする（S 1 0 3）。なお、ここで、差分データ（更新データ）を正ボリューム 1 8 にコピーするとは、正ボリューム 1 8 のうち障害ディスクドライブ 1 6 を除いた他の正常なディスクドライブ 1 6 及び予備ディスクドライブ 1 6（S P）の所定アドレスに更新データをコピーすることを意味する。これは後述する他の実施例でも同様である。

【0086】

D K A 1 2 は、更新データを正ボリューム 1 8 にコピーした後、この更新データに対応する差分ビットに「0」をセットする（S 1 0 4）。次に、D K A 1 2 は、フィードバックポインタが先頭アドレスを示しているか否かを検査する（S 1 0 5）。フィードバックポインタがワークボリューム 1 8 の先頭アドレスに達している場合は（S105：YES）、ワークボリューム 1 8 を全て検査してデータ移行が完了したときなので、D K A 1 2 は、処理を終了する。

【0087】

フィードバックポインタが先頭アドレスに達していない場合（S105：NO）、D K A 1 2 は、フィードバックポインタを 1 つ前に（古いデータに）移動させる（S 1 0 6）。そして、D K A 1 2 は、フィードバックポインタの示す更新データをジャーナルファイルから

読み出し、キャッシングメモリ14に記憶させる(S107)。DKA12は、キャッシングメモリ14に読み出した更新データに対応する差分ビットに「1」がセットされているか否かを判定する(S108)。差分ビットに「1」がセットされている場合(S108:YES)、DKA12は、この更新データを正ボリューム18にコピーし(S109)、差分ビットを「1」から「0」に変更し(S110)、S105に戻る。一方、キャッシングメモリ14に読み出した更新データに対応する差分ビットに「0」がセットされている場合(S108:NO)、DKA12は、キャッシングメモリ14に読み出したデータを削除し(S111)、S105に戻る。

【0088】

つまり、S105～S111では、新しいジャーナルファイルのデータから順番に正ボリューム18に反映させていき、古いジャーナルファイルのデータで新しいデータが上書きされるのを防止するために、差分ビットに「0」をセットする。そして、差分ビットに「0」がセットされたデータは、古いデータであると判断し、削除する。

【実施例3】

【0089】

図14～図18に基づいて、第3実施例を説明する。本実施例の1つの特徴は、データ退避領域として、ワークディスクを使用する点にある。

まず、図14は、本実施例によるデータ障害回避方法の全体動作の概略を示す説明図である。あるディスクドライブ16(#4)について障害の発生が予測されると、予備ディスクドライブ16(SP)へのデータ移行が開始される(S121)。

【0090】

このデータ移行の開始と共に、未使用のワークディスクドライブ16(W)が少なくとも1つ確保される(S122)。ホストコンピュータ1からの書き込み要求が発生すると、この更新データはワークディスクドライブ16(W)に記憶される(S123)。ワークディスクドライブ16(W)に記憶されたデータについては、差分管理テーブルT7により管理される(S124)。

【0091】

ホストコンピュータ1から読み出し要求が発行された場合、読み出すべきデータがデータ退避元である正のディスクドライブ16に存在するならば、正ディスクドライブ16からデータが読み出される(S125)。障害ディスクドライブ16(#4)に存在するデータを要求された場合、他の正常なディスクドライブ16(#1～3)の記憶内容に基づいてデータが復元され、復元されたデータがホストコンピュータ1に提供される。ホストコンピュータ1から要求されたデータがワークディスクドライブ16(W)に存在する場合、ワークディスクドライブ16(W)からデータが読み出され、ホストコンピュータ1に提供される(S126)。そして、予備ディスクドライブ16(SP)へのデータ移行が完了すると、ワークディスクドライブ16に退避されたデータが正ディスクドライブ16(障害ディスクドライブを除く)及び予備ディスクドライブ16(SP)に反映される(S127)。

【0092】

図15は、ディスクアレイ装置10内に記憶される各種管理テーブルの構造例を示す説明図である。図15(a)は、ディスク管理テーブルT5を示す。ディスク管理テーブルT5には、ディスクアレイ装置10の備える全てのディスクドライブ16について、ディスクドライブ番号と、記憶容量と、ステータスとが対応付けられている。ステータスとしては、少なくとも「更新データ退避中」と「NULL」とがある。図示の例では、正ディスクドライブ16(#1～4)がデータ退避モードに入っていることを示している。

【0093】

図15(b), (c)は、ワークディスク管理テーブルT6を示す。図15(b)は、予備ディスクドライブ16(SP)へのデータ移行前における状態を、図15(c)は、データ移行後の状態をそれぞれ示す。ワークディスク管理テーブルT6は、ワークディスクドライブ番号と、記憶容量と、ステータスと、対応する正ディスクドライブ番号と、更

新データを記憶する終端アドレスとを対応付けて管理する。

【0094】

データ移行前の状態では、2つのワークディスク16（#60, 61）は、いずれも「未使用」ステータスであり、正ディスクドライブ16に対応付けられていない。データ移行が開始されると、図示の例では、1つのワークディスクドライブ16（#60）が、4個の正ディスクドライブ16（#1～4）に対応付けられる。ステータスは「未使用」から「使用中」に変化する。1つのワークディスクドライブ16（#60）には、4つの正ディスクドライブ16（#1～4）を対象とする更新データがそれぞれ記憶され、最新の更新データの位置は終端アドレスとして示される。

【0095】

図16は、ディスクアレイ装置10内に記憶される差分管理テーブルT7を示す説明図である。差分管理テーブルT7は、「差分管理情報」の一例であって、正ディスクドライブ番号と、正ディスクドライブ16におけるアドレスと、ワークディスクドライブ番号と、ワークディスクドライブ16におけるアドレスとを対応付けている。図示の例では、正ディスク16（#1）のアドレス「1」、「2」に記憶されるべきデータが、ワークディスクドライブ16（#60）のアドレス「1」、「2」にそれぞれ退避していることが示されている。また、図示の例では、正ディスクドライブ16（#2）のアドレス「5」、「2」、「6」に記憶されるべきデータが、ワークディスクドライブ16（#60）のアドレス「3」、「4」、「5」にそれぞれ記憶されている。さらに、図示の例では、正ディスクドライブ16（#3）のアドレス「3」に記憶されるべきデータが、ワークディスクドライブ16（#60）のアドレス「6」に記憶されている。そして、ワークディスクドライブアドレス「6」の位置が終端アドレスとなっている。

【0096】

次に、図17は、DKA12により実行されるデータ退避処理を示すフローチャートである。S131～S135は、データ退避領域がディスクである点を除いて、図12で述べたS71～S75とほぼ同様である。即ち、データ移行が開始されると（S131：YES）、DKA12は、ワークディスクドライブ16が登録されているか否かを判定し（S132）、登録されているワークディスクドライブ16を順番に検査することにより（S133～4）、未使用のワークディスクドライブ16を検出する（S133：YES）。

【0097】

DKA12は、予備ディスクドライブ16（SP）へのデータ移行が完了するまでの間（S137）、ホストコンピュータ1からのアクセス要求が発生したか否かを監視する（S136）。データ移行が完了した場合（S137：YES）、ワークディスクドライブ16に退避させた更新データを正ディスクドライブ16及び予備ディスクドライブ16（SP）に反映させる（S138）。差分データのフィードバックが完了した後、ワークディスク管理テーブルT6から正ディスクドライブ番号等を削除し、ステータスを「未使用」に戻して、ワークディスクドライブ16を解放する（S139）。なお、データ移行中に、正ディスクドライブ16のステータスは「更新データ退避中」にセットされ、データ移行が終了すると、ステータスは「NULL」に変更される。

【0098】

データ移行中にホストコンピュータ1からのアクセス要求が発生すると（S136：YES）、DKA12は、要求されたデータが差分管理テーブルT7に登録されているか否かを判定する（S140）。要求されたデータが差分管理テーブルT7に登録されている場合（S140：YES）、DKA12は、ホストコンピュータ1からのアクセス要求が読み出し要求であるか否かを判定する（S141）。読み出し要求の場合（S141：YES）、DKA12は、ワークディスクドライブ16から目的のデータを読み出し（S142）、キャッシュメモリ14に記憶させ、S137に戻る。書き込み要求の場合（S141：NO）、DKA12は、更新データをワークディスクドライブ16に記憶させ（S143）、S137に戻る。ここで注意すべき点は、ジャーナルファイルとは異なり、同一アドレスに対する重複したデータ書き込みは、上書き処理される点である。

【0099】

ホストコンピュータ1から要求されたデータが差分管理テーブルT7に登録されていない場合(S140: NO)、DKA12は、ホストコンピュータ1からのアクセス要求が読み出し要求であるか否かを判定する(S144)。読み出し要求ではない場合(S144: NO)、DKA12は、更新データを記憶するだけの空き容量がワークディスクドライブ16に存在するか否かを判定する(S145)。ワークディスクドライブ16に残容量がある場合(S145: YES)、DKA12は、更新データの記憶先アドレス等を差分管理テーブルT7に登録する(S146)。また、DKA12は、終端アドレスを差分管理テーブルT7に登録し(S147)、ワークディスクドライブ16の終端アドレスに更新データを記憶させる(S148)。

【0100】

ワークディスクドライブ16に残容量が無い場合(S145: NO)、DKA12は、更新データを正ディスクドライブ16に記憶させて(S149)、S137に戻る。差分管理テーブルT7に登録されていないデータの読み出し要求である場合(S144: YES)、DKA12は、正ディスクドライブ16からデータを読み出し(S150)、キャッシュメモリ14に記憶させてS137に戻る。

【0101】

図18は、差分データのフィードバック処理を示すフローチャートである。本処理は、図17中のS138に対応する。DKA12は、差分管理テーブルT7にデータが登録されているか否かを判定する(S160)。差分管理テーブルT7にデータが登録されていない場合(S160: NO)、正ディスクドライブ16にフィードバックすべきデータが存在しないので、処理を終了する。

【0102】

差分管理テーブルT7にデータが登録されている場合(S160: YES)、DKA12は、差分管理テーブルT7に登録されたワークディスクアドレスに基づいて、ワークディスクドライブ16から全てのデータを読み出し、この読み出したデータをキャッシュメモリ14に記憶させる(S161)。DKA12は、キャッシュメモリ14に読み出した全データを、対応する正ディスクドライブ16の対応するアドレスにそれぞれコピーさせる(S162)。そして、DKA12は、差分管理テーブルT7を削除する(S163)。なお、図示の例では、ワークディスクドライブ16に退避させたデータの全てをキャッシュメモリ14に読み出す場合を説明したが、これに限らず、1アドレス分のデータずつキャッシュメモリ14に読み出して正ディスクドライブ16にコピーさせててもよい。

【実施例4】

【0103】

図19、図20に基づいて第4実施例を説明する。本実施例の1つの特徴は、複数のRAIDグループのそれぞれでスペアリング処理が実施された場合でも、対応できるようにした点にある。本実施例は、第2実施例及び第3実施例のいずれにも適用可能であるが、図19では、第2実施例の変形例として説明する。

【0104】

本実施例では、RAIDグループ17(P1)とRAIDグループ17(P2)との複数のRAIDグループにおいて、それぞれ独自にディスクドライブ16の障害発生が予測される。そして、障害ディスクドライブ16が検出されると、それぞれ別々の予備ディスクドライブ16(SP1), (SP2)に障害ディスクドライブ16のデータがコピーされる(S171)。

【0105】

いずれか1つのRAIDグループ17においてスペアリング処理が開始されると、登録されたワークボリュームのうち空いているワークボリューム18(S)が確保され、データ退避元の論理ボリューム18と対応付けられる(S172)。また、別のRAIDグループにおいてスペアリング処理が開始されると、別のワークボリューム18(S)が確保される。図示の例では、第1のRAIDグループ17(P1)の論理ボリューム18(P

1) は、ワークボリューム18 (S1) に対応し、第2のRAIDグループ17 (P2) の論理ボリューム18 (P2) は、ワークボリューム18 (S1) に対応する。

【0106】

データ移行中に、ホストコンピュータ1から書き込み要求があった場合は、対応するワークボリューム18 (S) にデータが書き込まれる。差分ビットマップ20は、ワークボリューム18 (S) に登録されたデータを管理する (S174)。

【0107】

データ移行中に、ホストコンピュータから読み出し要求があった場合、要求されたデータが正の論理ボリューム18に存在するときは、正の論理ボリューム18からデータが復元されて、ホストコンピュータ1に提供される (S175)。ホストコンピュータ1から要求されたデータがワークボリューム18 (S) に存在するときは、ワークボリューム18 (S) からデータが読み出される (S176)。

【0108】

データ移行が完了すると、ワークボリューム18 (S) に退避させていたデータを正の論理ボリューム18及び予備ディスクドライブ16 (SP) にそれぞれ反映させる (S177)。なお、以上の各処理は、各RAIDグループそれぞれについて、独立して実行される。

【0109】

図20は、ディスクアレイ装置10に記憶される管理テーブルを示す。図20 (a) は、第2実施例と同様のワークボリューム管理テーブルT4を示す。図11 (b) に示す第2実施例のワークボリューム管理テーブルとの相違点は、各ワークボリューム18 (S) に複数の正ボリュームを対応付けることが可能となっている点である。

【0110】

例えば、本実施例のワークボリューム管理テーブルT4では、ワークボリューム18 (#10) に、2つの正ボリューム18 (#1, 4) が対応付けられている。例えば、一方の正ボリューム18 (#1) はRAIDグループ17 (P1) に属し、他方の正ボリューム18 (#4) は他のRAIDグループ17 (P2) に属する。このように、ワークボリューム18 (S) は、それぞれ異なるRAIDグループ17の論理ボリューム18に対応付け可能である。

【0111】

図20 (b) は、第3実施例に適用した場合におけるワークディスク管理テーブルT6を示す。図15 (c) に示すワークディスク管理テーブルとの相違点は、1つのワークディスクドライブ16 (#60) に、複数のRAIDグループをそれぞれ構成する複数の正ディスクドライブ16 (#1~8) を対応付け可能である点である。

【0112】

このように、各管理テーブルT4, T6を拡張するだけで、データ障害回避処理を多重起動させることができる。

【実施例5】

【0113】

図21~図24に基づいて第5実施例を説明する。本実施例の1つの特徴は、障害発生が予測されたディスクドライブ16への書き込み要求のみを退避させる点にある。本実施例は第1実施例~第3実施例にそれぞれ適用可能であるが、図21では、第2実施例に適用した場合を例に挙げて説明する。

【0114】

図21は、データ障害回避方法の全体概要を示す説明図である。障害ディスクドライブ16が検出されて予備ディスクドライブ16 (SP) へのデータ移行が開始すると (S181)、空いているワークボリューム18 (S) が確保される (S182)。このワークボリューム18 (S) は、障害ディスクドライブ16 (#4) に書き込まれるべきデータ(パーティティを含む)を退避するために用いられる。ワークボリューム18 (S) には、他の正常なディスクドライブ16 (#1~3) を対象とするデータは書き込まれない。

【0115】

データ移行中に、ホストコンピュータ1から障害ディスクドライブ16(#4)を対象とする書込み要求が発行された場合、更新データは、ワークボリューム18(S)に記憶される(S183)。差分ビットマップ20は、ワークボリューム18(S)に記憶されたデータについて管理する(S184)。

【0116】

データ移行中に、ホストコンピュータ1から読み出し要求があった場合、要求されたデータが正常なディスクドライブ16(#1~3)に存在するならば、正ディスクドライブ16から目的のデータが読み出される(S185)。読み出しを要求されたデータが障害ディスクドライブ16(#4)に存在すべき場合は、ワークボリューム18(S)からデータが読み出される(S186)。

【0117】

一方、データ移行中に、ホストコンピュータ1から正常なディスクドライブ16(#1~3)を対象とする書込み要求があった場合、それぞれのディスクドライブ16(#1~3)に対してデータが書き込まれる(S187)。

【0118】

そして、データ移行が完了すると、ワークボリューム18(S)に退避させたデータは、予備ディスクドライブ16(SP)にコピーされる(S188)。

【0119】

図22は、本実施例を第1実施例に適用した場合におけるデータ退避処理を示すフローチャートである。この実施例では、障害ディスクドライブ16のデータ退避領域として、RAIDグループを使用する。

【0120】

本処理のS191~S197は、図8で述べたS31~S37と同一の処理を行うので説明を省略する。ホストコンピュータ1からのアクセス要求が発生すると(S194:YES)、DKA12は、アクセスを要求されたデータ(パリティを含む)が障害ディスクドライブ16に存在するか否かを判別する(S198)。障害ディスクドライブ16以外のディスクドライブ16に存在するデータを要求された場合(S198:NO)、DKA12は、ホストコンピュータ1からのアクセス要求が読み出し要求であるか否かを判定する(S199)。書込み要求の場合(S199:NO)、DKA12は、正ボリューム(正ディスクドライブ。本処理において以下同様)にデータを書き込み(S200)、S195に戻る。ホストコンピュータ1からのアクセス要求が読み出し要求の場合(S199:YES)、DKA12は、正ボリュームからデータを読み出す(S201)。

【0121】

障害ディスクドライブ16を対象とするアクセス要求の場合(S198:YES)、DKA12は、このアクセス要求が読み出し要求であるか否かを判定する(S202)。読み出し要求の場合(S202:YES)、DKA12は、要求されたデータに対応する差分ビットに「1」がセットされているか否かを判定する(S203)。差分ビットに「0」がセットされている場合(S203:NO)、データは更新されていないので、DKA12は、要求されたデータを正ボリュームのデータに基づいて復元し(S201)、S195に戻る。差分ビットに「1」がセットされている場合(S203:YES)、更新済のデータなので、DKA12は、ワークボリューム(副ボリュームである。本処理において以下同様)18からデータを読み出し(S204)、S195に戻る。

【0122】

障害ディスクドライブ16を対象とするアクセス要求であって、かつ書込み要求である場合(S202:NO)、DKA12は、対応する差分ビットに「1」をセットし(S205)、ワークボリューム18にデータを書き込んで(S206)、S195に戻る。

【0123】

図23は、本実施例を第2実施例に適用した場合のデータ退避処理を示すフローチャートである。本処理のS211~S219は、図12で述べたS71~S79と同一の処理

を行うので、説明を割愛する。

【0124】

DKA12は、予備ディスクドライブ16(S P)へのデータ移行中に、ホストコンピュータ1からのアクセス要求が発生すると(S216: YES)、要求されたデータが障害ディスクドライブ16に存在するか否かを判定する(S220)。正常な他のディスクドライブ16に存在するデータを対象とする場合(S220: YES)、DKA12は、ホストコンピュータ1からのアクセス要求が読み出し要求であるか否かを判定する(S221)。読み出し要求の場合(S221: YES)、DKA12は、正ボリュームからデータを読み出し(S222)、S217に戻る。書き込み要求の場合(S221: NO)、DKA12は、更新データを正ボリュームに書き込む(S223)。

【0125】

ホストコンピュータ1からのアクセス要求が障害ディスクドライブ16を対象とする場合(S220: YES)、DKA12は、このアクセス要求が読み出し要求であるか否かを判定する(S224)。読み出し要求の場合(S224: YES)、DKA12は、要求されたデータに対応する差分ビットに「1」がセットされているか否かを検査する(S225)。差分ビットに「1」がセットされている場合(S225: YES)、DKA12は、終端アドレスから上に向けて(古い方に向けて)目的のデータを検索する(S226)。そして、DKA12は、発見されたデータをワークボリューム18から読み出して(S227)、S217に戻る。要求されたデータに対応する差分ビットに「0」がセットされている場合(S225: NO)、DKA12は、正ボリュームからデータを読み出して(S228)、S217に戻る。

【0126】

ホストコンピュータ1からのアクセス要求が障害ディスクドライブ16を対象とする書き込み要求である場合(S224: NO)、DKA12は、ワークボリューム18に残容量があるか否かを検査する(S229)。ワークボリューム18に残容量が無い場合(S229: NO)、DKA12は、更新データを正ボリュームに書き込む(S230)。そして、DKA12は、更新データに対応する差分ビットに「0」をセットし(S231)、S217に戻る。ワークボリューム18に残量がある場合(S229: YES)、DKA12は、更新データに対応する差分ビットに「1」をセットし(S232)、更新データをワークボリューム18に書き込む(S233)。DKA12は、終端アドレスを更新して(S234)、S217に戻る。

【0127】

図24は、本実施例を第3実施例に適用した場合のデータ退避処理を示すフローチャートである。本処理のS241～S249は、図17で述べたS131～S139と同一の処理を行うので説明を省略する。

【0128】

予備ディスクドライブ16(S P)へのデータ移行中に、ホストコンピュータ1から障害ディスクドライブ16以外の正常なディスクドライブ16を対象とするアクセス要求が出された場合(S250: NO)、DKA12は、このアクセス要求が読み出し要求であるか否かを判別する(S251)。読み出し要求の場合(S251: YES)、DKA12は、正ディスクドライブ16からデータを読み出し(S252)、S247に戻る。書き込み要求の場合(S251: NO)、DKA12は、更新データを正ディスクドライブ16に書き込み(S253)、S247に戻る。

【0129】

一方、ホストコンピュータ1から障害ディスクドライブ16を対象とするアクセス要求が出された場合(S250: YES)、DKA12は、差分管理テーブルT7に登録されているデータが要求されているか否かを判定する(S254)。差分管理テーブルT7に登録されているデータの場合(S254: YES)、DKA12は、ホストコンピュータ1からのアクセス要求が読み出し要求であるか否かを判定する(S255)。読み出し要求の場合(S255: YES)、DKA12は、ワークディスクからデータを読み出し(S256)、S247に

戻る。

【0130】

差分管理テーブルT7に登録されていないデータを対象とするアクセス要求の場合(S254: NO)、DKA12は、このアクセス要求が読み出し要求であるか否かを判定する(S258)。書き込み要求の場合(S258: NO)、DKA12は、ワークディスクに残容量があるか否かを検査する(S259)。ワークディスクに残容量がある場合(S259: YES)、DKA12は、更新データの記憶元アドレス等を差分管理テーブルT7に登録する(S260)。また、DKA12は、終端アドレスを差分管理テーブルT7に登録し(S261)、ワークディスクに更新データを書き込んで(S262)、S247に戻る。

【実施例6】

【0131】

図25～図29に基づいて、第6実施例を説明する。本実施例の1つの特徴は、スペアリング処理及びデータ退避処理のいずれにおいても、正常なディスクドライブに記憶されたデータに基づいて、障害ディスクドライブに記憶されたデータを復元し、この復元したデータを予備ディスクドライブにコピーさせると共に、ホストコンピュータに提供するようにした点にある。

【0132】

本実施例は、第1実施例～第3実施例にそれぞれ適用可能であるが、図25では、第1実施例に適用した場合を例に挙げて説明する。図25は、本実施例によるデータ障害回避方法の全体動作の概要を示す説明図である。

【0133】

前記各実施例と同様に、RAIDグループ17(P)を構成するディスクドライブ16(#4)に障害の発生が予測されると、予備ディスクドライブ16(SP)へのデータ移行が開始される(S271)。ここで、注意すべき点は、障害ディスクドライブ16(#4)から直接データを読み出して予備ディスクドライブ16(SP)にコピーするのではなく、他の正常なディスクドライブ16(#1～3)の記憶内容に基づいて障害ディスクドライブ16(#4)内のデータを復元し、この復元したデータを予備ディスクドライブ16(SP)にコピーする点である。従って、スペアリング処理中に、障害ディスクドライブ16(#4)からの読み出しは行われない。

【0134】

予備ディスクドライブ16(SP)へのデータ移行が開始されると、未使用のRAIDグループ17(S)が確保され(S272)、正のRAIDグループ17(P)とペアを形成する(S273)。また、ここで、副RAIDグループ17(S)には、正RAIDグループ17(P)の正論理ボリューム18(P)に対応する副ボリューム18(S)が形成される。

【0135】

データ移行中に、ホストコンピュータ1から正のRAIDグループ17(P)を対象とする書き込み要求が発行された場合、この更新データは、副ボリューム18(S)に記憶される(S274)。副ボリューム18(S)に記憶されたデータについては、差分ビットマップ20により管理される(S275)。

【0136】

データ移行中に、ホストコンピュータ1から、更新されていないデータの読み出し要求が出された場合は、正ボリューム18(P)からデータを読み出して、ホストコンピュータ1に提供する(S276)。障害ディスクドライブ16(#4)に記憶されているデータの読み出し要求の場合は、他の正常なディスクドライブ16(#1～3)からのデータに基づいてデータを復元する。

【0137】

データ移行中に、ホストコンピュータ1から、更新済データの読み出し要求が出された場合は、副ボリューム18(S)からデータを読み出して、ホストコンピュータ1に提供する(S277)。そして、データ移行が終了すると、差分ビットマップ20に基づいて、

副ボリューム18（S）の記憶内容を正ボリューム18（P）（障害ディスクドライブを除く）及び予備ディスクドライブ16（SP）に反映させる（S278）。

【0138】

図26は、本実施例によるスペアリング処理（データ移行処理）を示すフローチャートである。まず、DKA12は、コピー元ディスクドライブ（障害ディスクドライブ）の先頭アドレスにセットする（S281）。DKA12は、コピー元ディスクドライブ以外の他の正常なディスクドライブ16から、コピー元ディスクドライブ16のデータをキャッシュメモリ14にコピーする（S282）。

【0139】

データ復元に使用するストライプデータのキャッシュメモリ14への読み出しが正常に終了した場合（S283:YES）、DKA12は、キャッシュメモリ14に読み出されたデータに基づいて逆演算を行い、コピー元ディスクドライブに存在するはずのデータを復元する（S284）。データの復元が正常に終了した場合（S285:YES）、DKA12は、復元したデータを予備ディスクドライブ16（SP）に書き込む（S286）。予備ディスクドライブ16（SP）へのデータ書き込みが正常に終了した場合（S287:YES）、DKA12は、コピー元ディスクドライブの終端アドレスに達したか否か、即ち、データ移行を全て完了したか否かを判定する（S288）。データ移行が完了していない場合（S288:N0）、DKA12は、コピー元ディスクドライブの終端アドレスに達したか否かを判定する（S289）。データ移行が完了するまでの間、S282～S289の処理が繰り返される。

【0140】

正常なディスクドライブ16からストライプデータの読み出しに失敗した場合（S283:N0）、DKA12は、目的とするデータをコピー元ディスクドライブ16から直接読み出し、キャッシュメモリ14に記憶させる（S291）。コピー元ディスクドライブ16からのデータ読み出しに成功した場合（S292:YES）、S286に移る。コピー元ディスクドライブ16からのデータ読み出しに失敗した場合（S292:N0）、コピー対象のデータは消失されたものとして扱い（S293）、S288に移る。

【0141】

一方、復元されたデータを予備ディスクドライブ16（SP）へ正常に書き込みできなかった場合（S287:N0）、対象データの書き込みエラーとして扱い（S290）、S288に移る。

【0142】

図27は、本実施例を第1実施例に適用した場合のデータ退避処理を示す。本処理の多くのステップは、図22で述べたステップと同一の処理を実行する。そこで、S314を中心に述べる。障害ディスクドライブ16を対象とする読み出し要求が発行され（S312:YES）、この要求されたデータが更新されていない場合（S313:N0）、DKA12は、他の正常なディスクドライブ16から読み出したデータに基づいて、目的のデータを復元し（S314）、S305に戻る。

【0143】

図28は、本実施例を第2実施例に適用した場合のデータ退避処理を示す。本処理の多くのステップは、図23で述べたステップと同一の処理を実行する。そこで、S338を中心に説明する。障害ディスクドライブ16を対象とする読み出し要求が発行され（S334:YES）、この要求されたデータが更新されていない場合（S335:N0）、DKA12は、他の正常なディスクドライブ16から読み出したデータに基づいて、目的のデータを復元し（S338）、S327に戻る。

【0144】

図29は、本実施例を第3実施例に適用した場合のデータ退避処理を示す。前記同様に、本処理の多くのステップは、図24で述べたステップと同一の処理を実行する。図24と異なるステップは、S374である。S374において、DKA12は、障害ディスクドライブ16以外の正常なディスクドライブ16からデータを読み出し、目的とするデータ

タを復元する（S374）。

【0145】

なお、本発明は、上述した各実施の形態に限定されない。当業者であれば、本発明の範囲内で、種々の追加や変更等を行うことができる。例えば、実施例中で明示した組合せ以外でも各実施例を適宜組合せ可能である。

【図面の簡単な説明】

【0146】

【図1】本発明の実施例に係わるディスクアレイ装置の全体概要を示すブロック図である。

【図2】RAID構成管理テーブルの構成を示す説明図であって、（a）はスペアリング処理の実行前、（b）はスペアリング処理の実行後の状態をそれぞれ示す。

【図3】ペア情報管理テーブルの構成を示す説明図であって、（a）はスペアリング処理の実行前、（b）はスペアリング処理の実行後の状態をそれぞれ示す。

【図4】差分ビットマップの構成を示す説明図である。

【図5】第1実施例におけるデータ障害回避方法の全体概要を示す説明図である。

【図6】スペアリング処理を示すフローチャートである。

【図7】手動でスペアリング処理を行う場合のフローチャートである。

【図8】データ退避処理を示すフローチャートである。

【図9】差分データのフィードバック処理を示すフローチャートである。

【図10】第2実施例におけるデータ障害回避方法の全体概要を示す説明図である。

【図11】ワークボリューム管理テーブル等を示す説明図であって、（a）はスペアリング処理の実行前、（b）はスペアリング処理の実行後の状態、（c）はワークボリュームの記憶構造をそれぞれ示す。

【図12】データ退避処理を示すフローチャートである。

【図13】差分データのフィードバック処理を示すフローチャートである。

【図14】第3実施例におけるデータ障害回避方法の全体概要を示す説明図である。

【図15】管理テーブルを示す説明図であって、（a）はディスク管理テーブル、（b）はスペアリング処理実行前のワークディスク管理テーブル、（c）はスペアリング処理実行後のワークディスク管理テーブルをそれぞれ示す。

【図16】差分管理テーブルを示す説明図である。

【図17】データ退避処理を示すフローチャートである。

【図18】差分データのフィードバック処理を示すフローチャートである。

【図19】第4実施例におけるデータ障害回避方法の全体概要を示す説明図である。

【図20】（a）はワークボリューム管理テーブル、（b）はワークディスク管理テーブルをそれぞれ拡張した様子を示す説明図である。

【図21】第5実施例におけるデータ障害回避方法の全体概要を示す説明図である。

【図22】データ退避処理を示すフローチャートである。

【図23】データ退避処理の別の例を示すフローチャートである。

【図24】データ退避処理のさらに別の例を示すフローチャートである。

【図25】第6実施例におけるデータ障害回避方法の全体概要を示す説明図である。

【図26】スペアリング処理を示すフローチャートである。

【図27】データ退避処理を示すフローチャートである。

【図28】データ退避処理の別の例を示すフローチャートである。

【図29】データ退避処理のさらに別の例を示すフローチャートである。

【符号の説明】

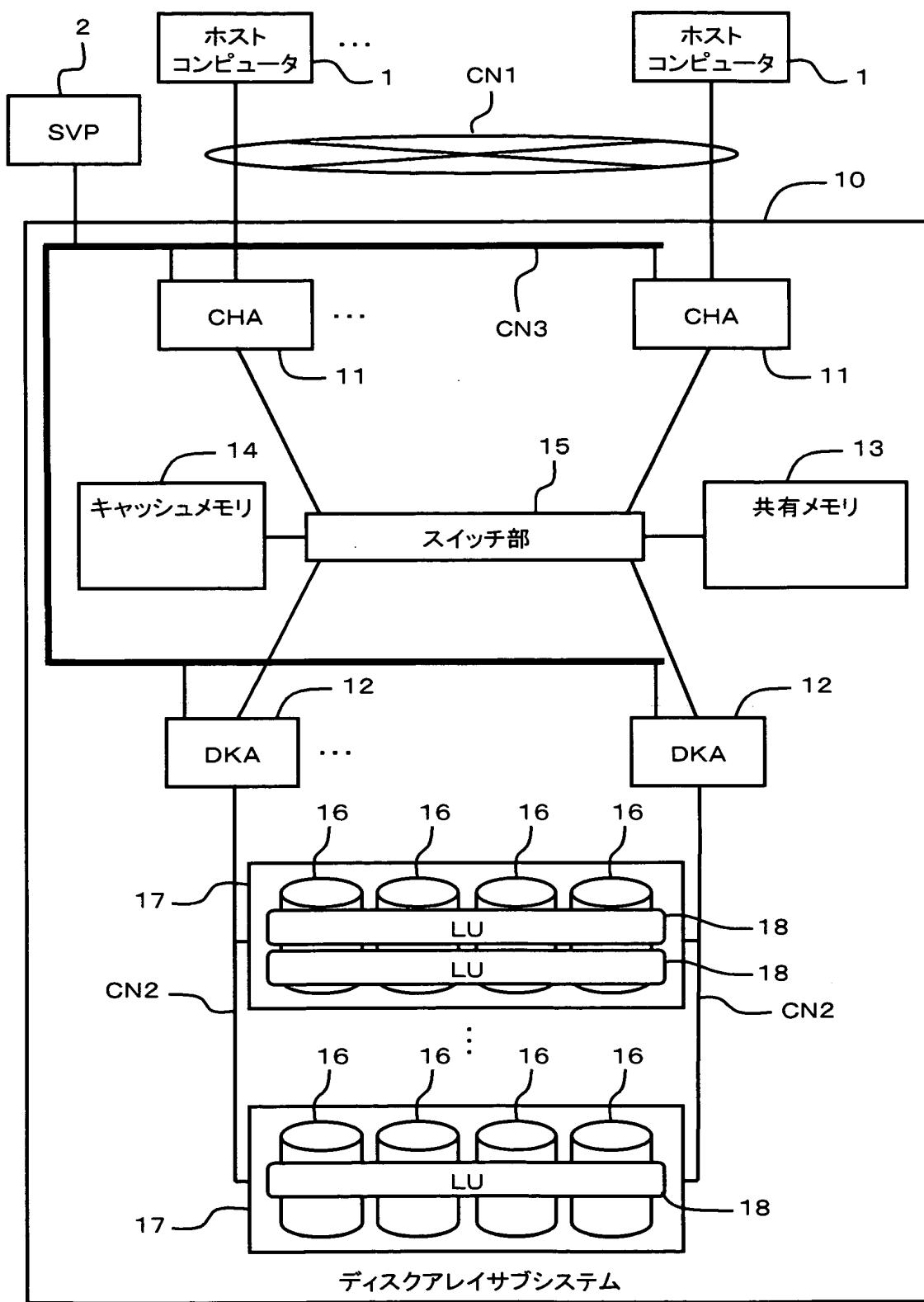
【0147】

1…ホストコンピュータ、2…SVP、10…ディスクアレイ装置、11…チャネルアダプタ、12…ディスクアダプタ、13…共有メモリ、14…キャッシュメモリ、15…スイッチ部、16…ディスクドライブ、17…RAIDグループ、18…論理ボリューム、20…差分ビットマップ、CN1～CN3…通信ネットワーク、T1…RAID構成管

理テーブル、T 2…ペア情報管理テーブル、T 3…エラー管理テーブル、T 4…ワークボリューム管理テーブル、T 5…ディスク管理テーブル、T 6…ワークディスク管理テーブル、T 7…差分管理テーブル、T h…閾値

【書類名】図面

【図 1】



【図2】

(a) RAID構成管理テーブル(変更前)

グループ#	ボリューム#	ディスク#	RAIDレベル
1	1, 2, 3	1, 2, 3, 4	RAID5
2	4, 5, 6	5, 6, 7, 8	RAID5
3	7, 8, 9, 10	9, 10, 11, 12	RAID1
4	11, 12	13, 14, 15, 16	RAID1
5	NULL	17, 18, 19, 20	RAID5
:	:	:	:

T1

(b) RAID構成管理テーブル(変更後)

グループ#	ボリューム#	ディスク#	RAIDレベル
1	1, 2, 3	1, 2, 3, 4	RAID5
2	4, 5, 6	5, 6, 7, 8	RAID5
3	7, 8, 9, 10	9, 10, 11, 12	RAID1
4	11, 12	13, 14, 15, 16	RAID1
5	13, 14, 15	17, 18, 19, 20	RAID5
:	:	:	:

T1

【図 3】

(a)ペア情報管理テーブル(変更前)

正ボリューム#	副ボリューム#	ペア状態	差分ビットマップ
4	7	二重化	010000100000…
5	8	二重化	100000010000…

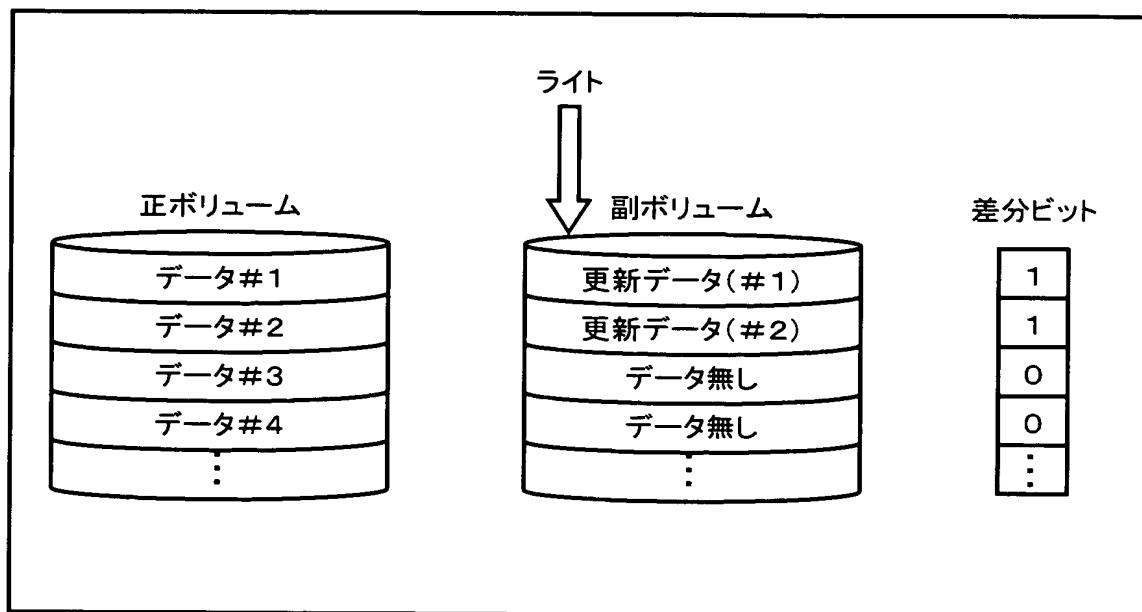
T2

(b)ペア情報管理テーブル(変更後)

正ボリューム#	副ボリューム#	ペア状態	差分ビットマップ
4	7	二重化	010000100000…
5	8	二重化	100000010000…
1	13	更新データ退避中	110000100011…
2	14	更新データ退避中	100001000000…
3	15	更新データ退避中	000000000000…

【図 4】

(a)

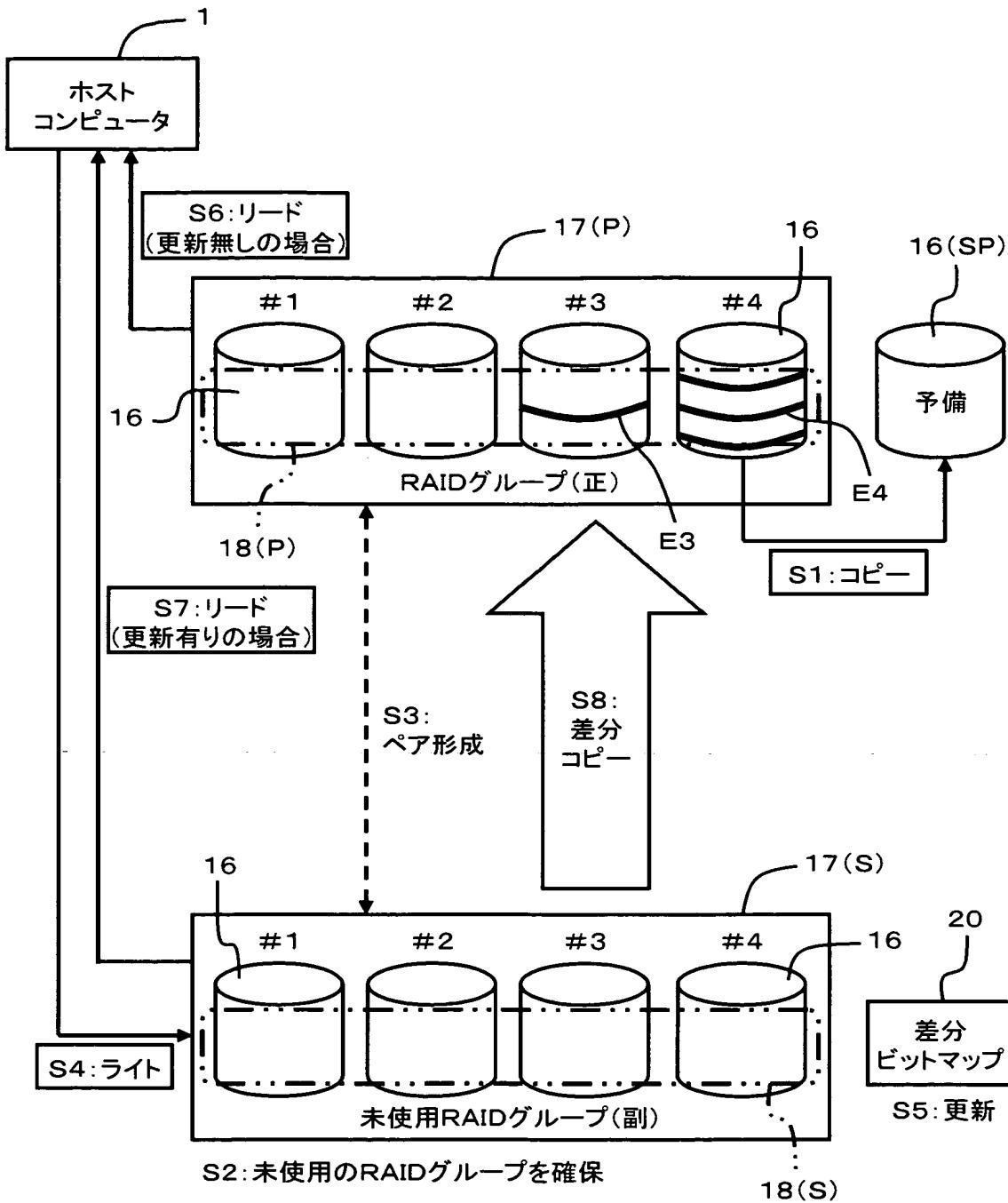


(b) 差分ビットマップ

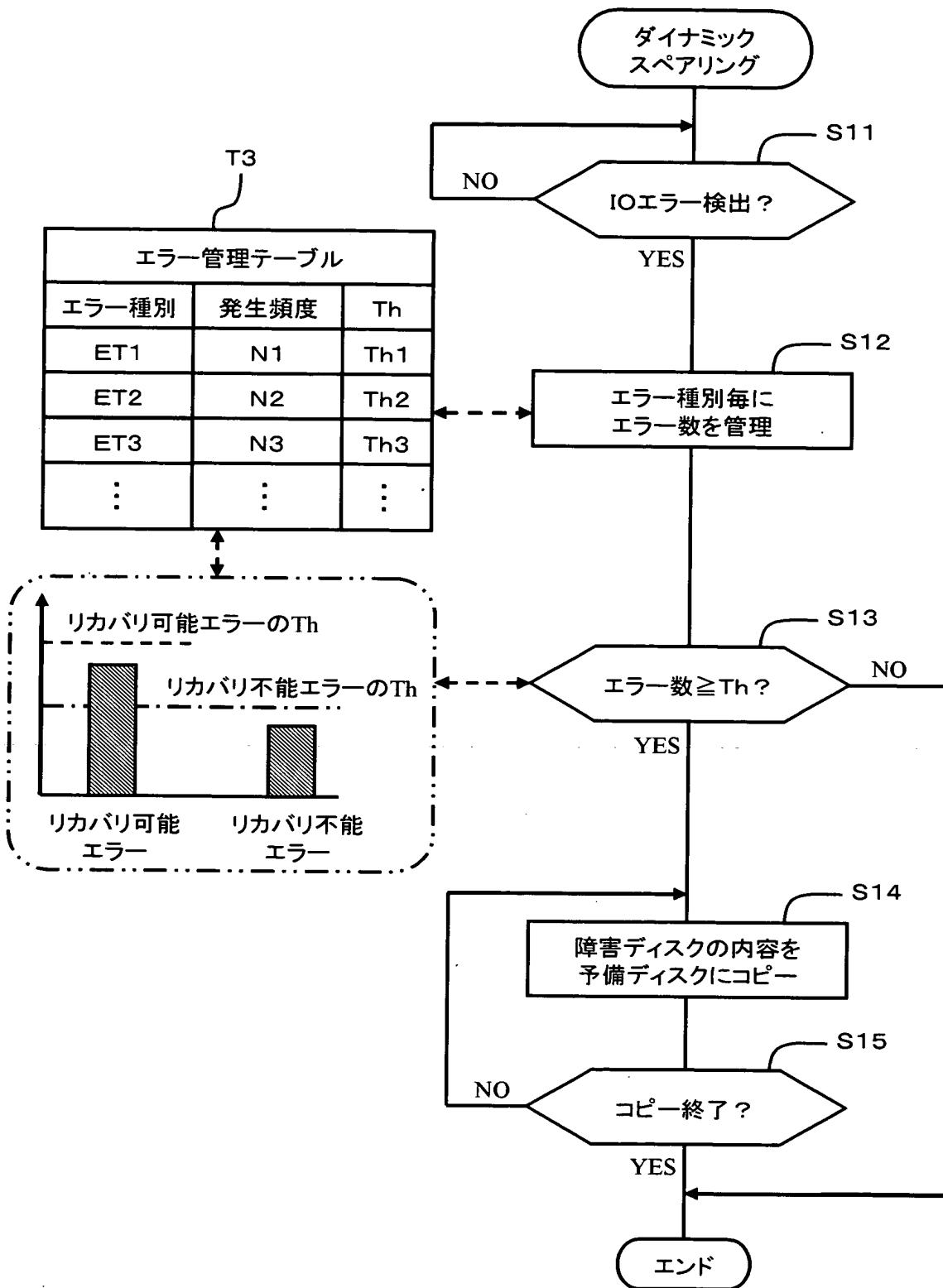
1100001000111100...

20

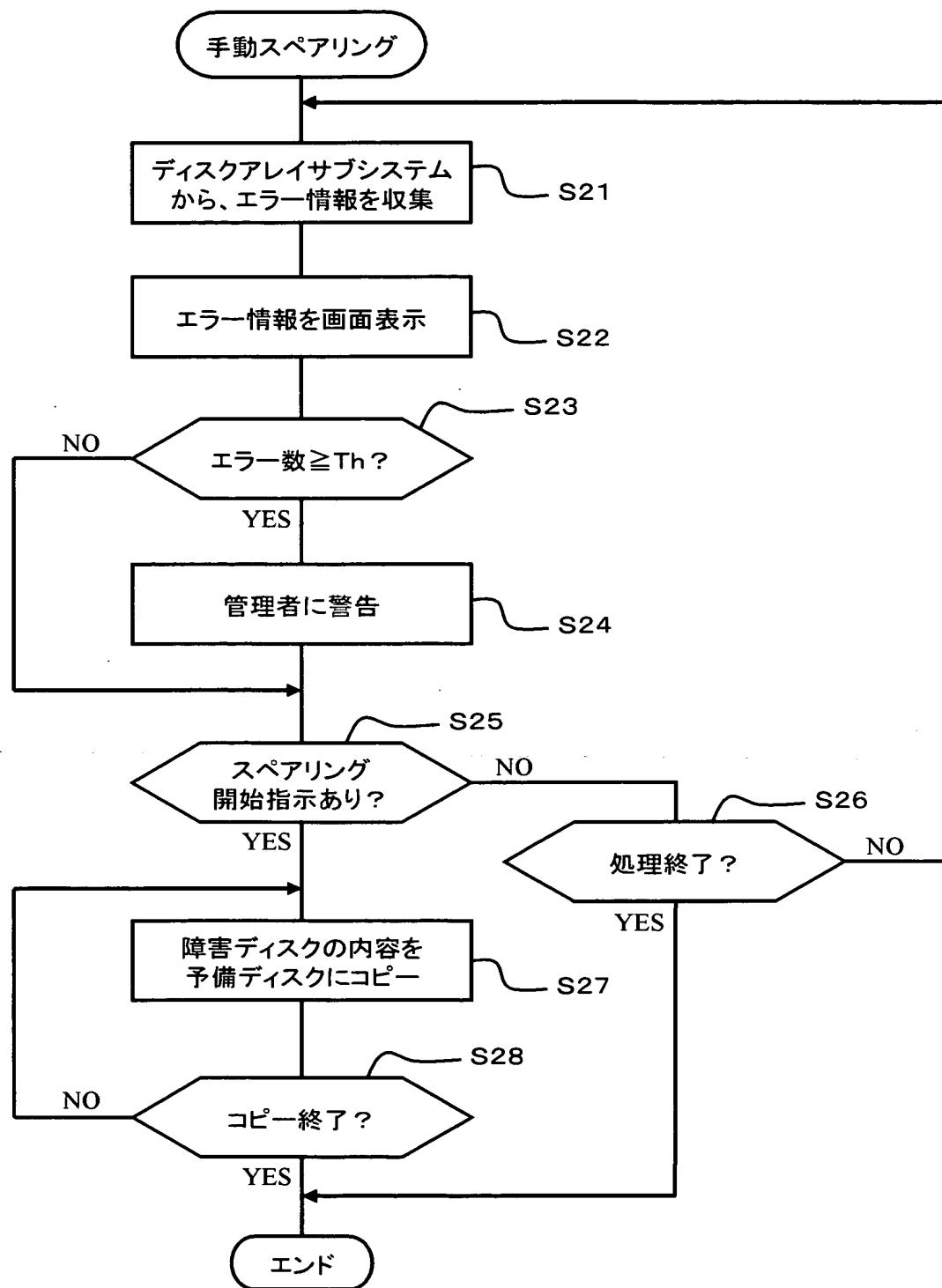
【図 5】



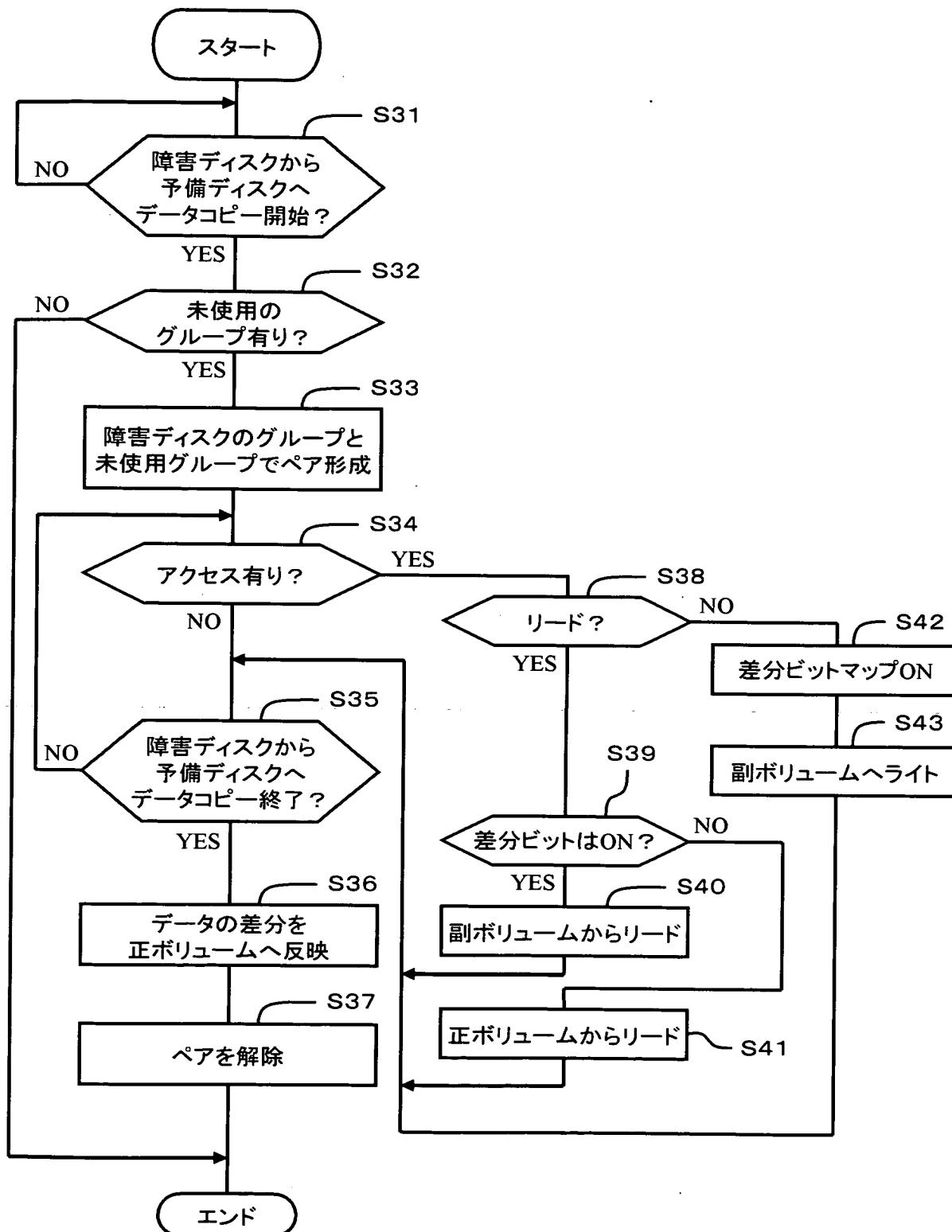
【図6】



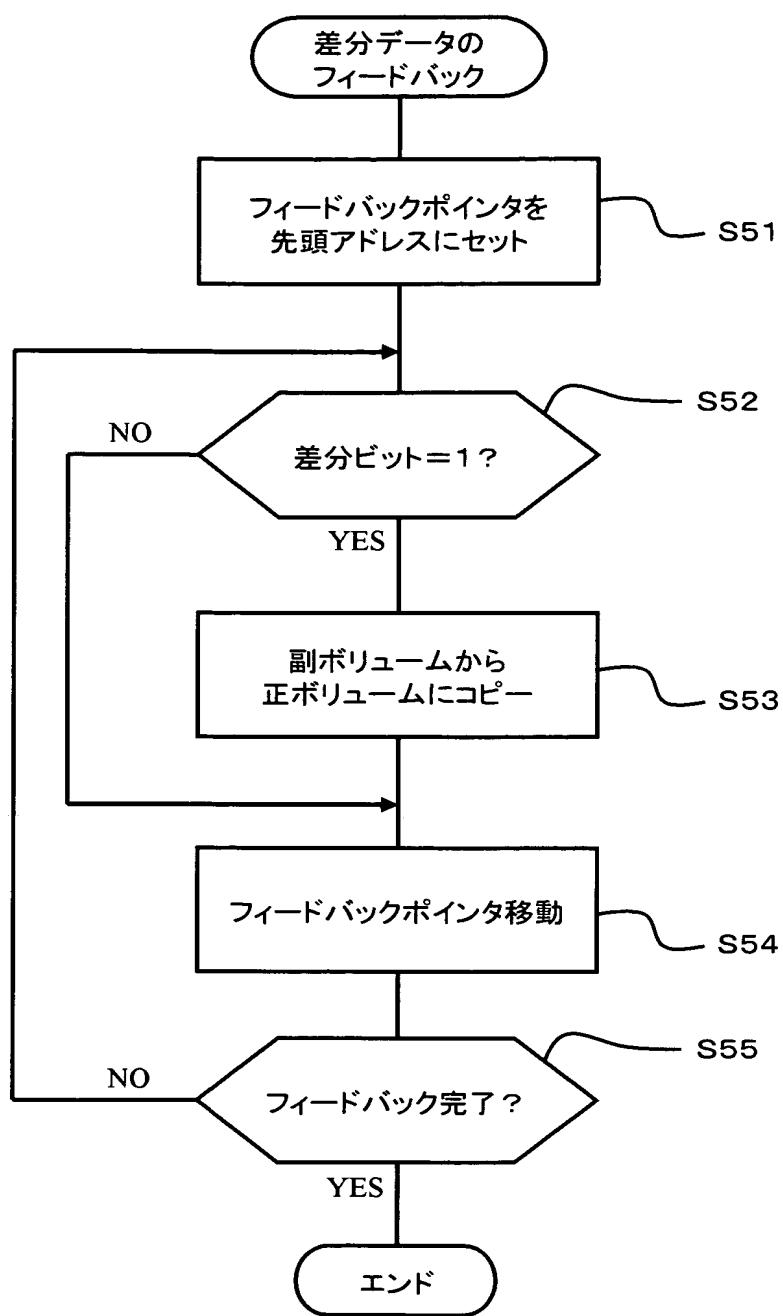
【図7】



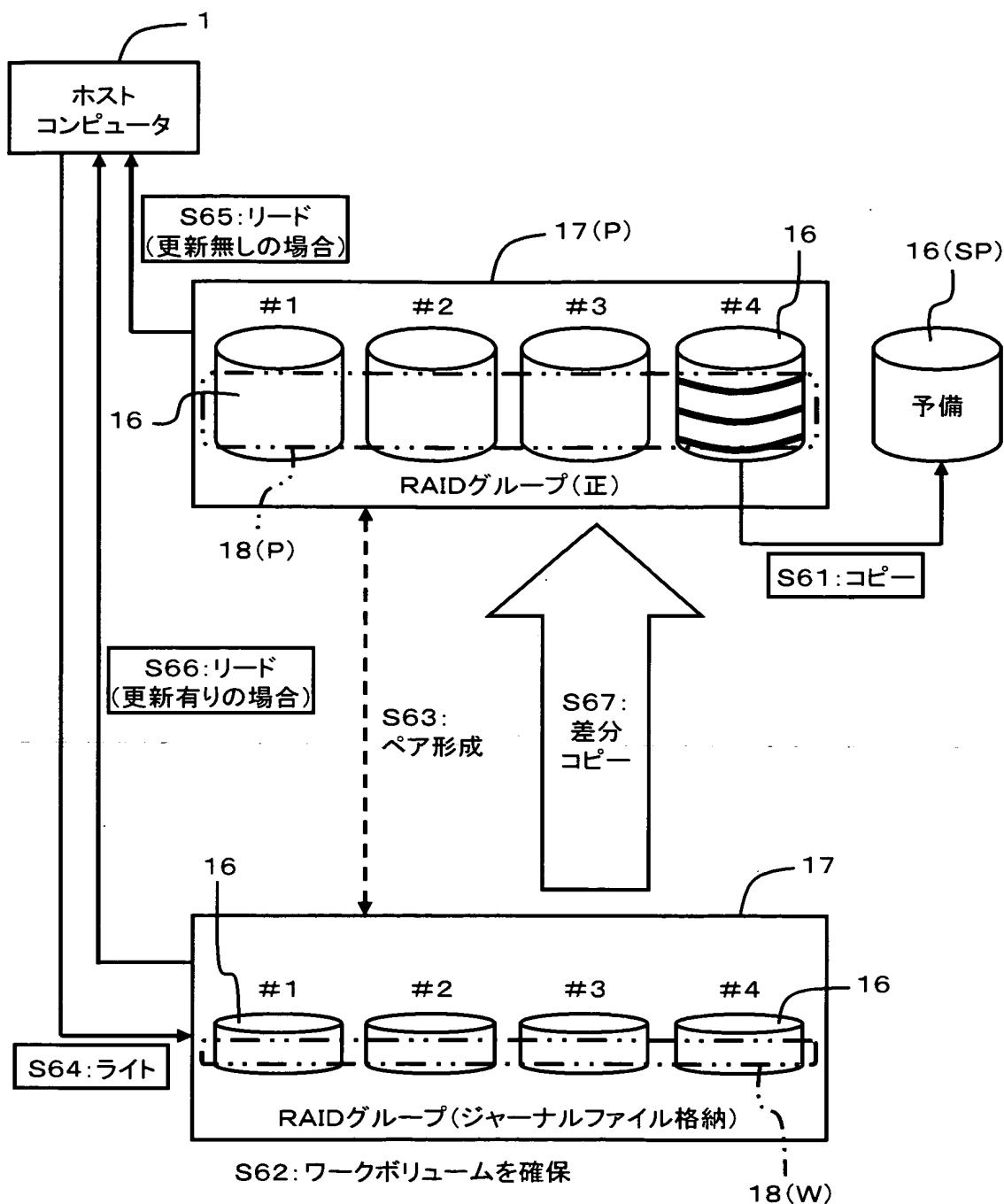
【図8】



【図9】



【図 10】



【図 1-1】

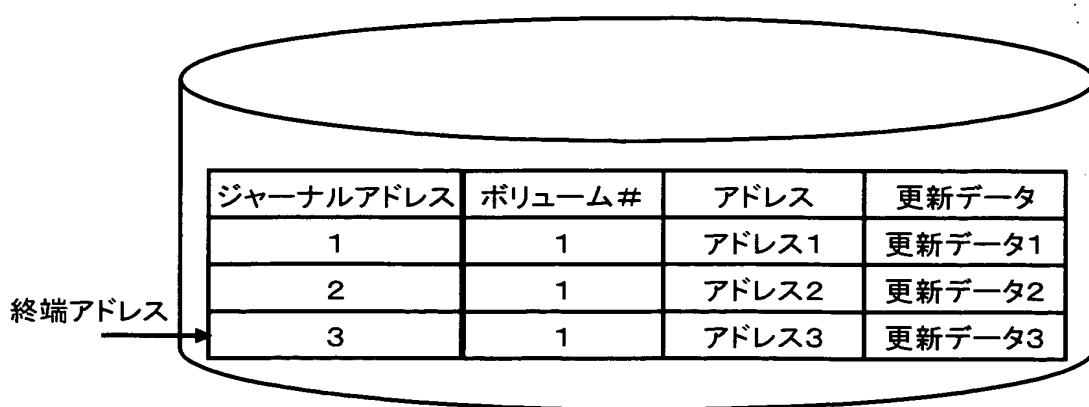
(a) ワークボリューム管理テーブル(スペアリング前)

ワークボリューム#	容量	正ボリューム#	終端アドレス	差分ビットマップ
10	3GB	NULL	NULL	000000000000…
11	3GB	NULL	NULL	000000000000…
12	3GB	NULL	NULL	000000000000…

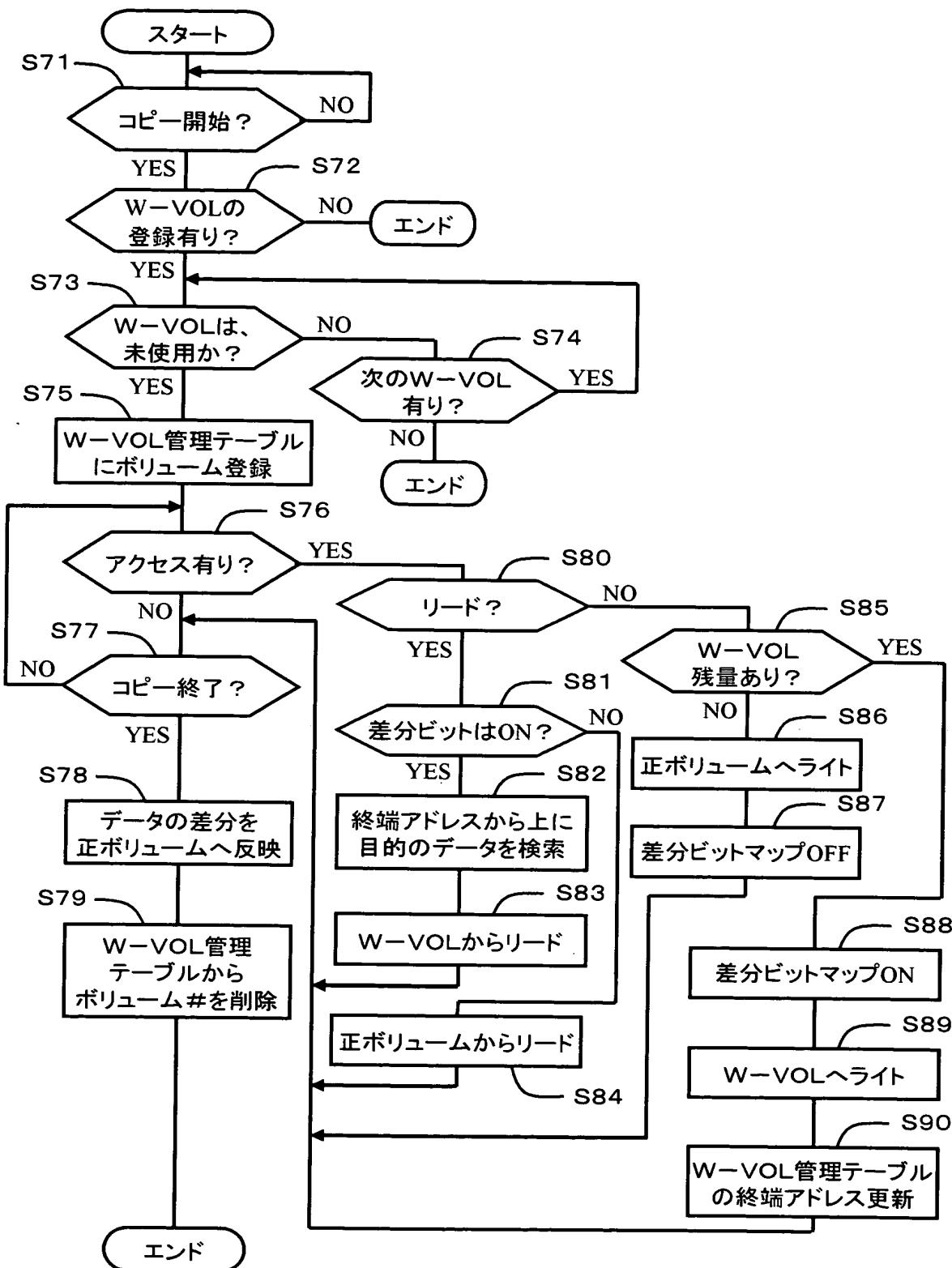
(b) ワークボリューム管理テーブル(スペアリング後)

ワークボリューム#	容量	正ボリューム#	終端アドレス	差分ビットマップ
10	3GB	1	3	1110000000…
11	3GB	2	NULL	0000000000…
12	3GB	3	NULL	0000000000…

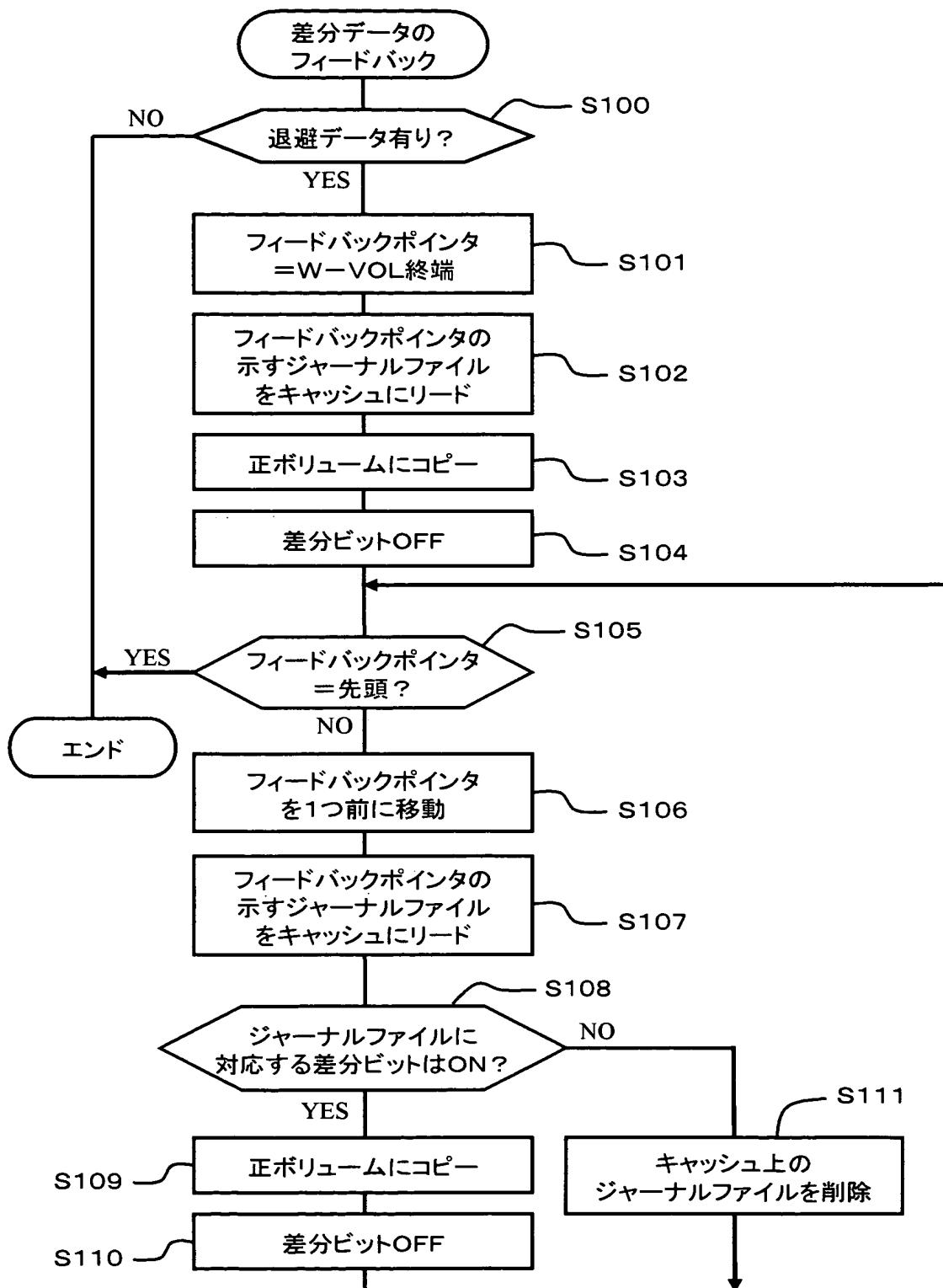
(c) ワークボリュームの記憶構造



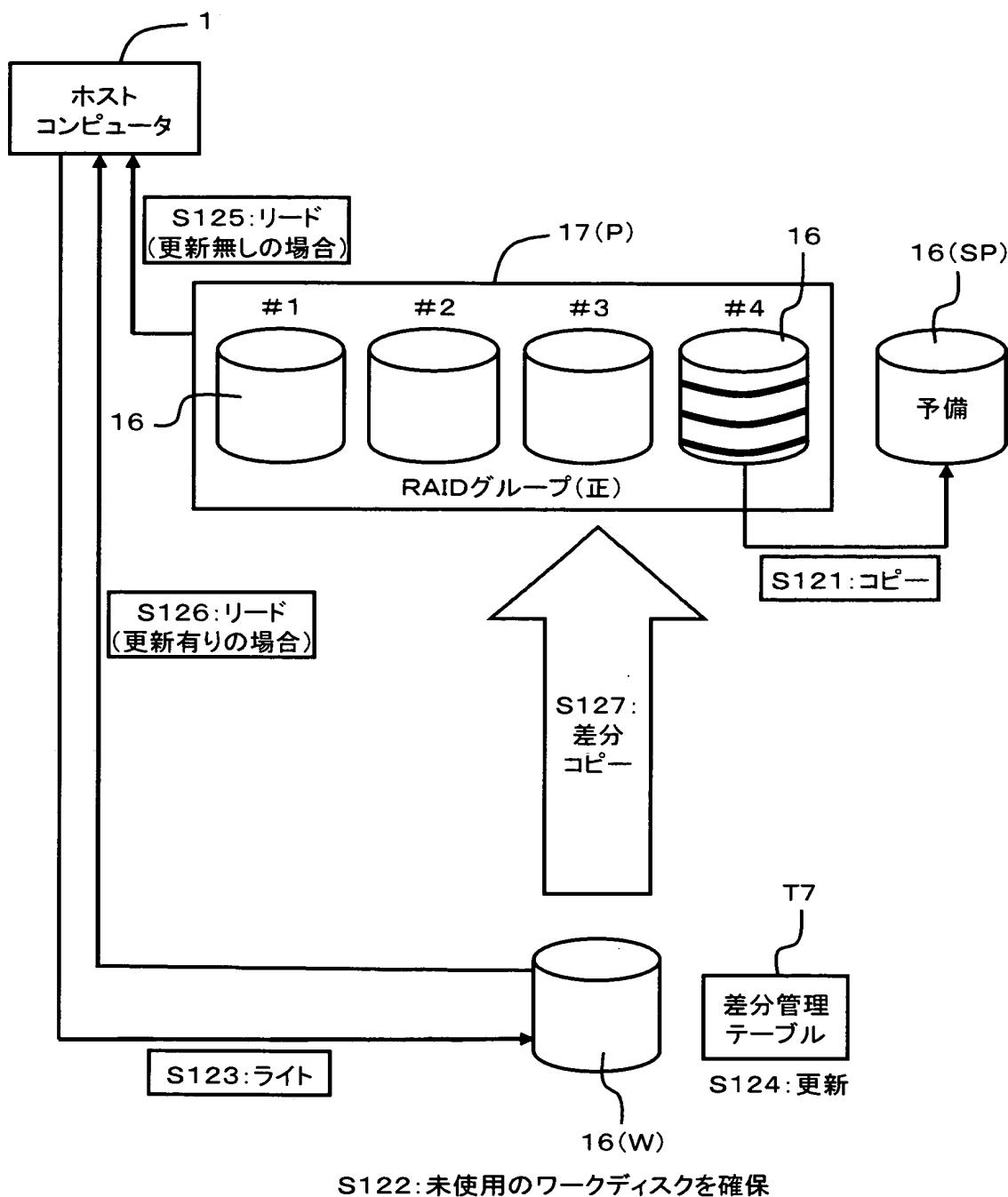
【図12】



【図13】



【図 14】



【図 15】

(a) ディスク管理テーブル

ディスク#	容量	ステータス
1	18GB	更新データ退避中
2	18GB	更新データ退避中
3	18GB	更新データ退避中
4	18GB	更新データ退避中
5	18GB	NULL
⋮	⋮	⋮

T5

(b) ワークディスク管理テーブル(スペアリング前)

ディスク#	容量	ステータス	正ディスク#	終端アドレス
60	18GB	未使用	NULL	NULL
61	18GB	未使用	NULL	NULL

T6

(c) ワークディスク管理テーブル(スペアリング後)

ディスク#	容量	ステータス	正ディスク#	終端アドレス
60	18GB	使用中	1, 2, 3, 4	6
61	18GB	未使用	NULL	NULL

T6

【図16】

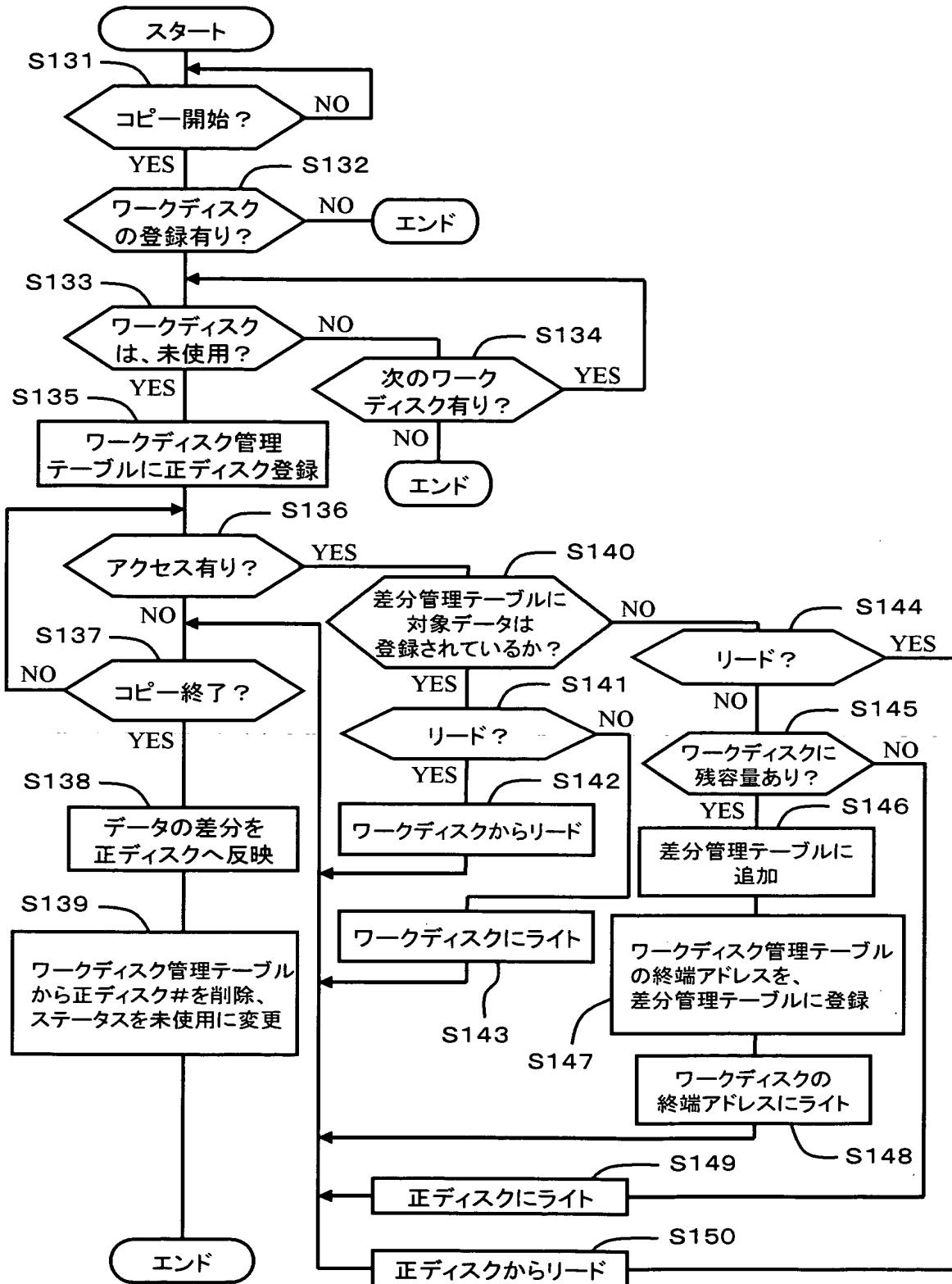
差分管理テーブル

T7

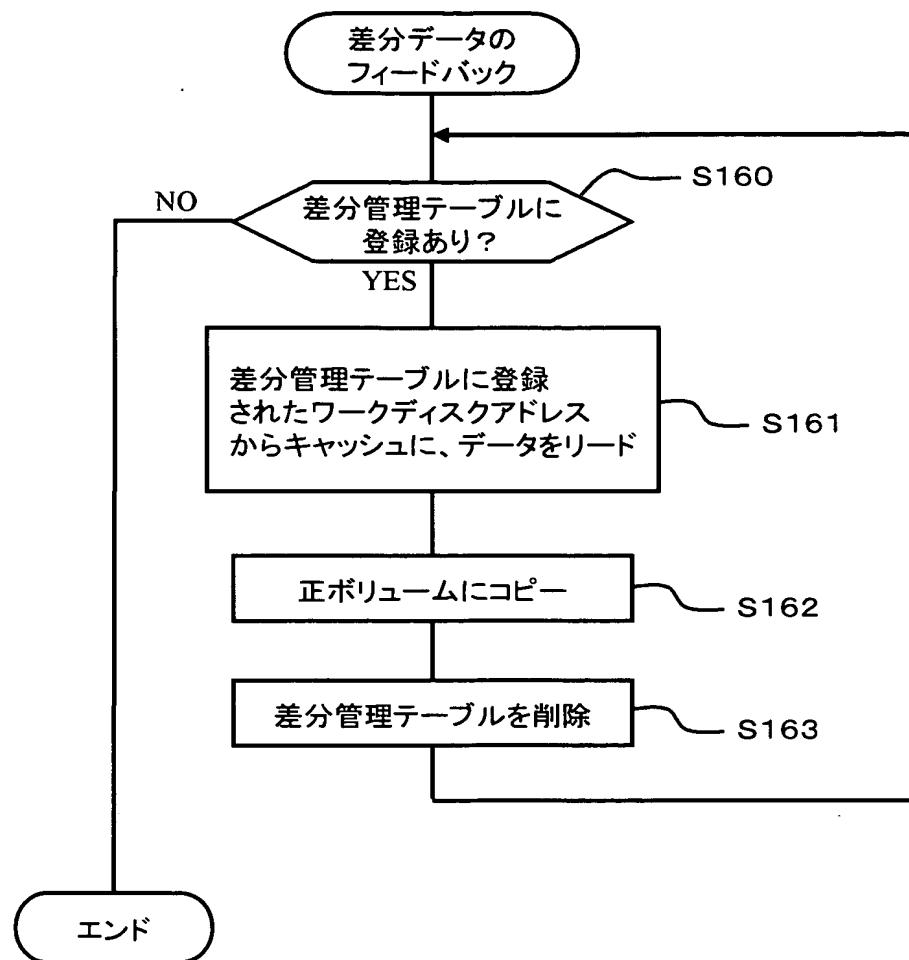
正ディスク#	正ディスク アドレス	ワークディスク#	ワークディスク アドレス
1	1	60	1
1	2	60	2
2	5	60	3
2	2	60	4
2	6	60	5
3	3	60	6

終端アドレス →

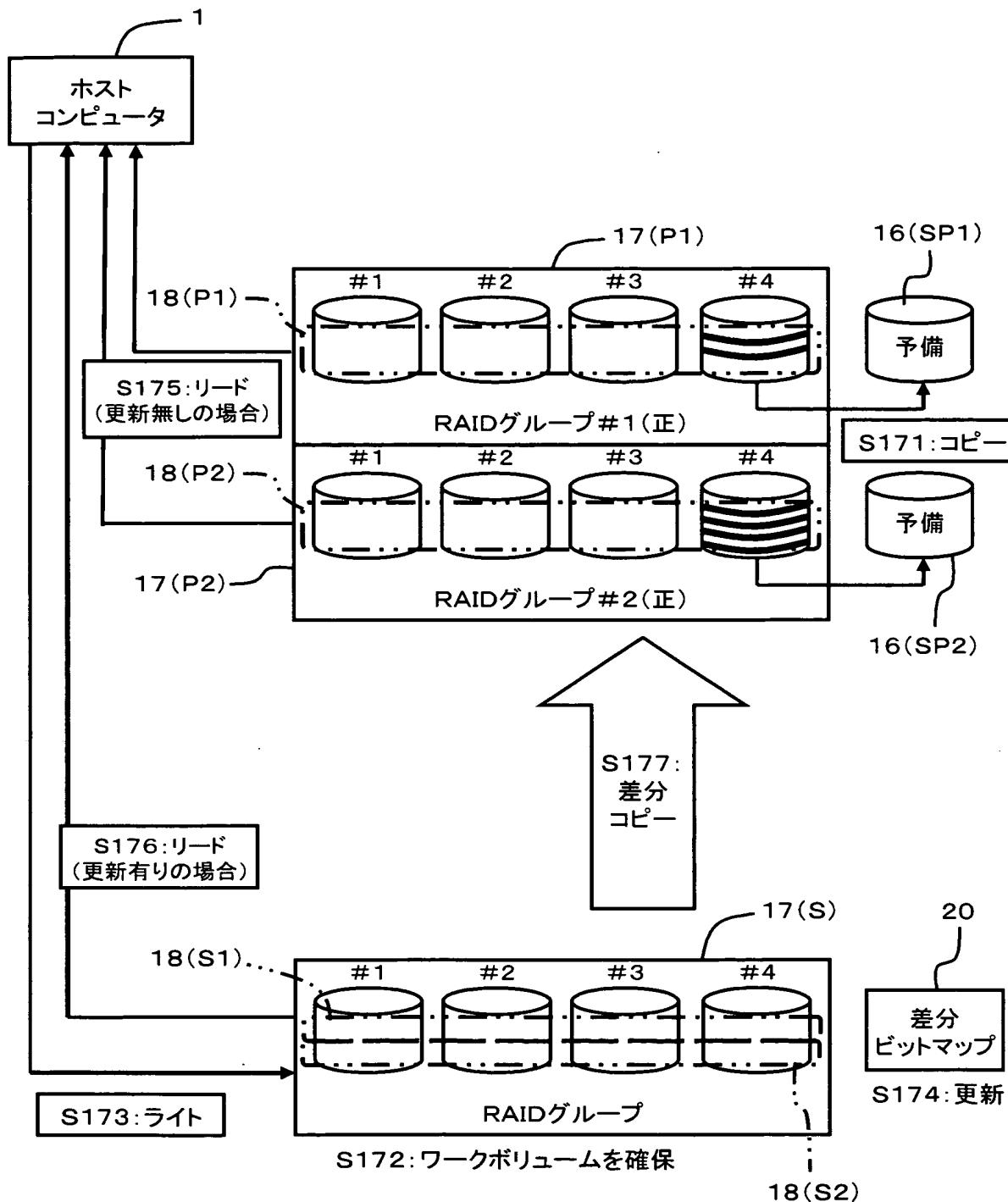
【図 17】



【図18】



【図19】



【図 20】

(a) ワークボリューム管理テーブル(スペアリング後)

ワークボリューム#	容量	正ボリューム#	終端アドレス	差分ビットマップ
10	3GB	1, 4	3	1110000000…
11	3GB	2, 5	NULL	0000000000…
12	3GB	3, 6	NULL	0000000000…

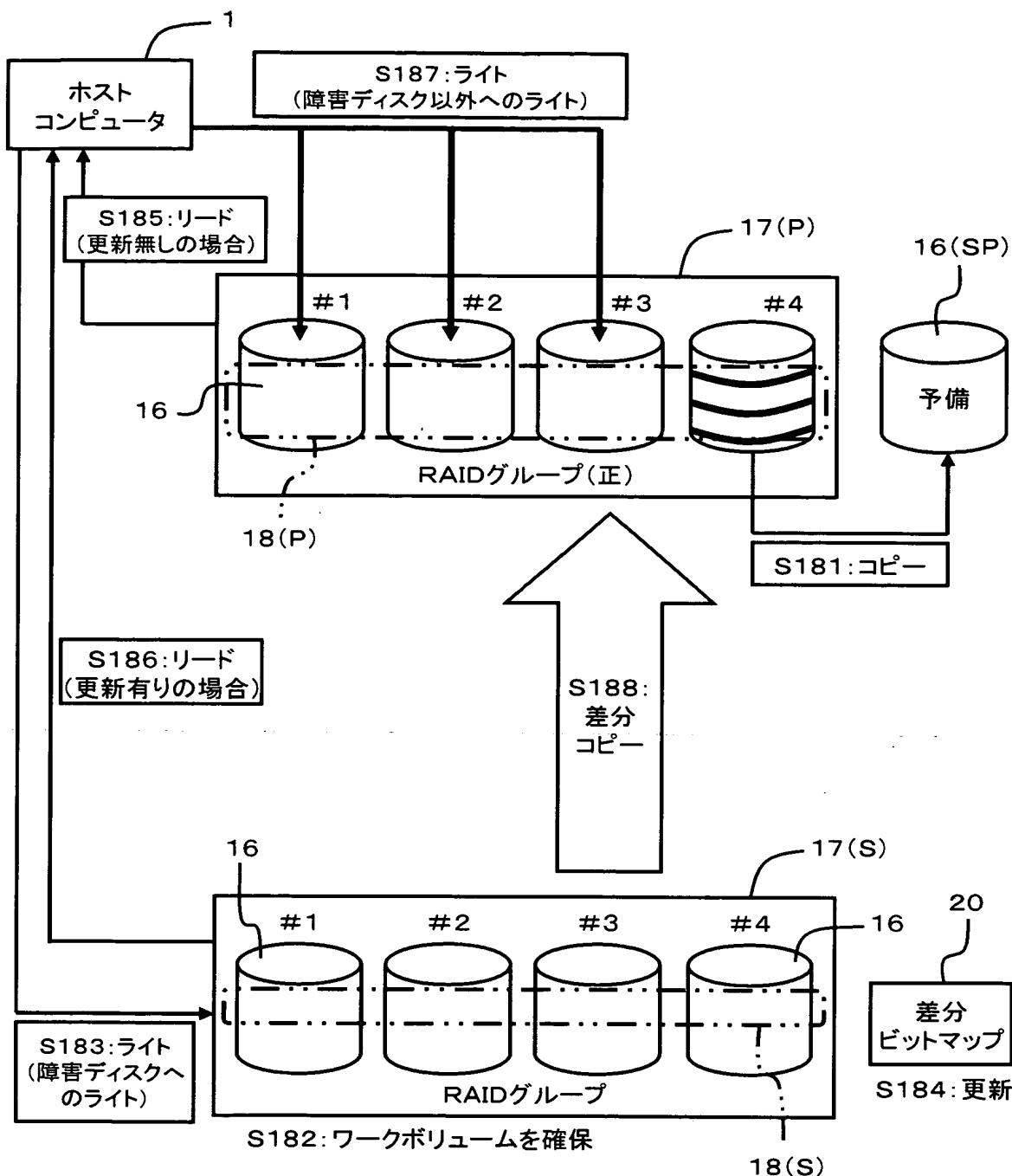
T4

(b) ワークディスク管理テーブル(スペアリング後)

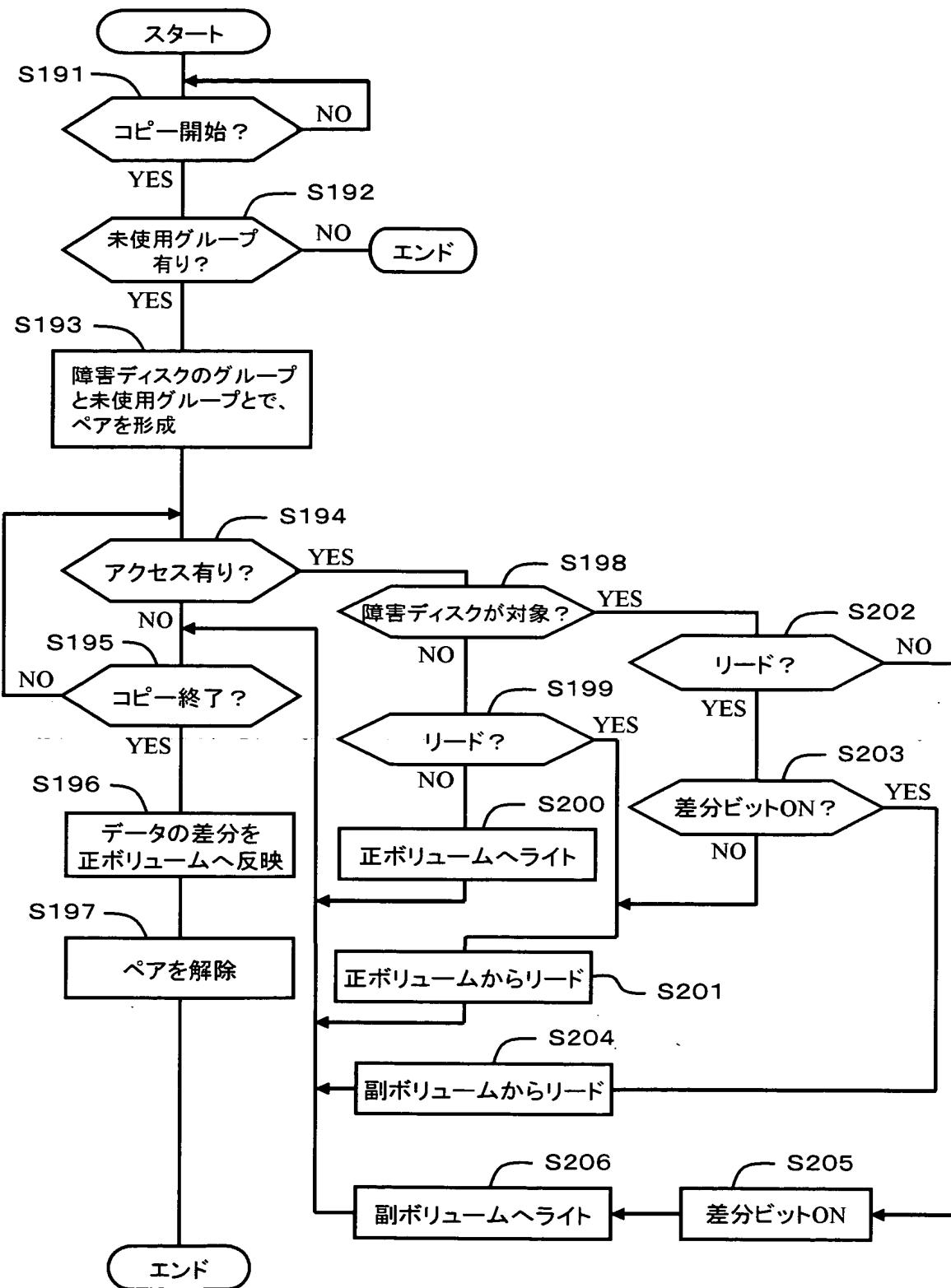
ディスク#	容量	ステータス	正ディスク#	終端アドレス
60	18GB	使用中	1, 2, 3, 4, 5, 6, 7, 8	6
61	18GB	未使用	NULL	NULL

T6

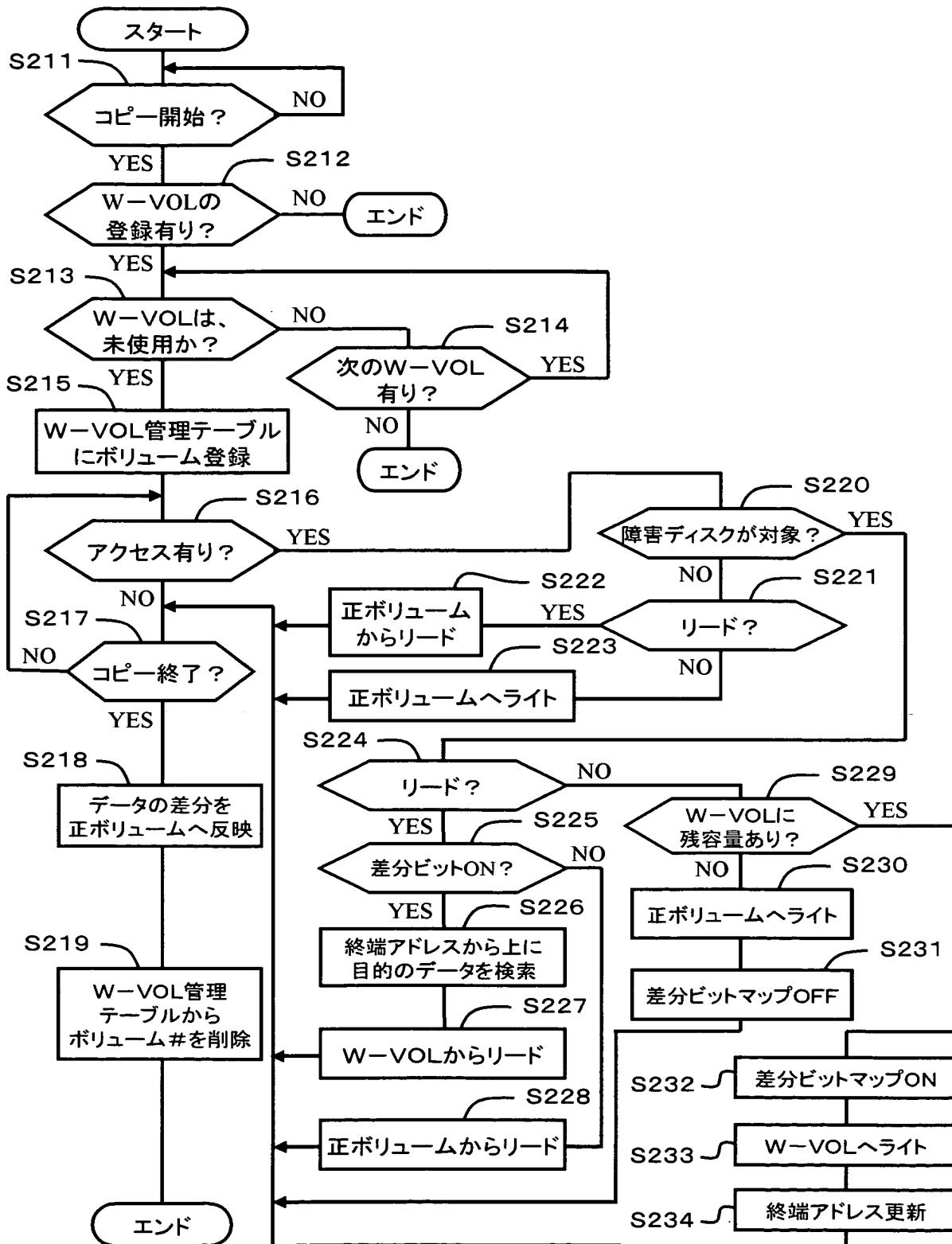
【図21】



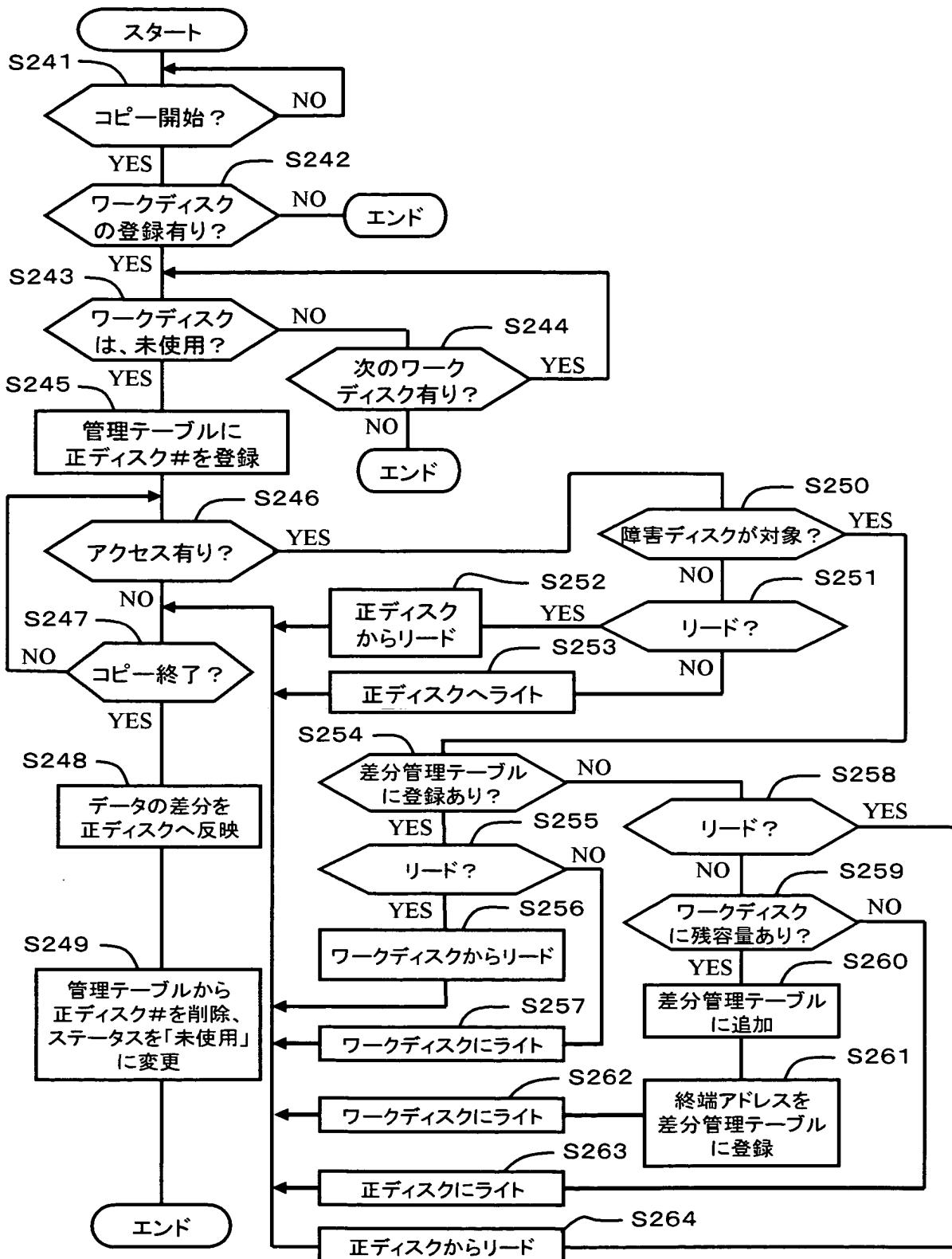
【図22】



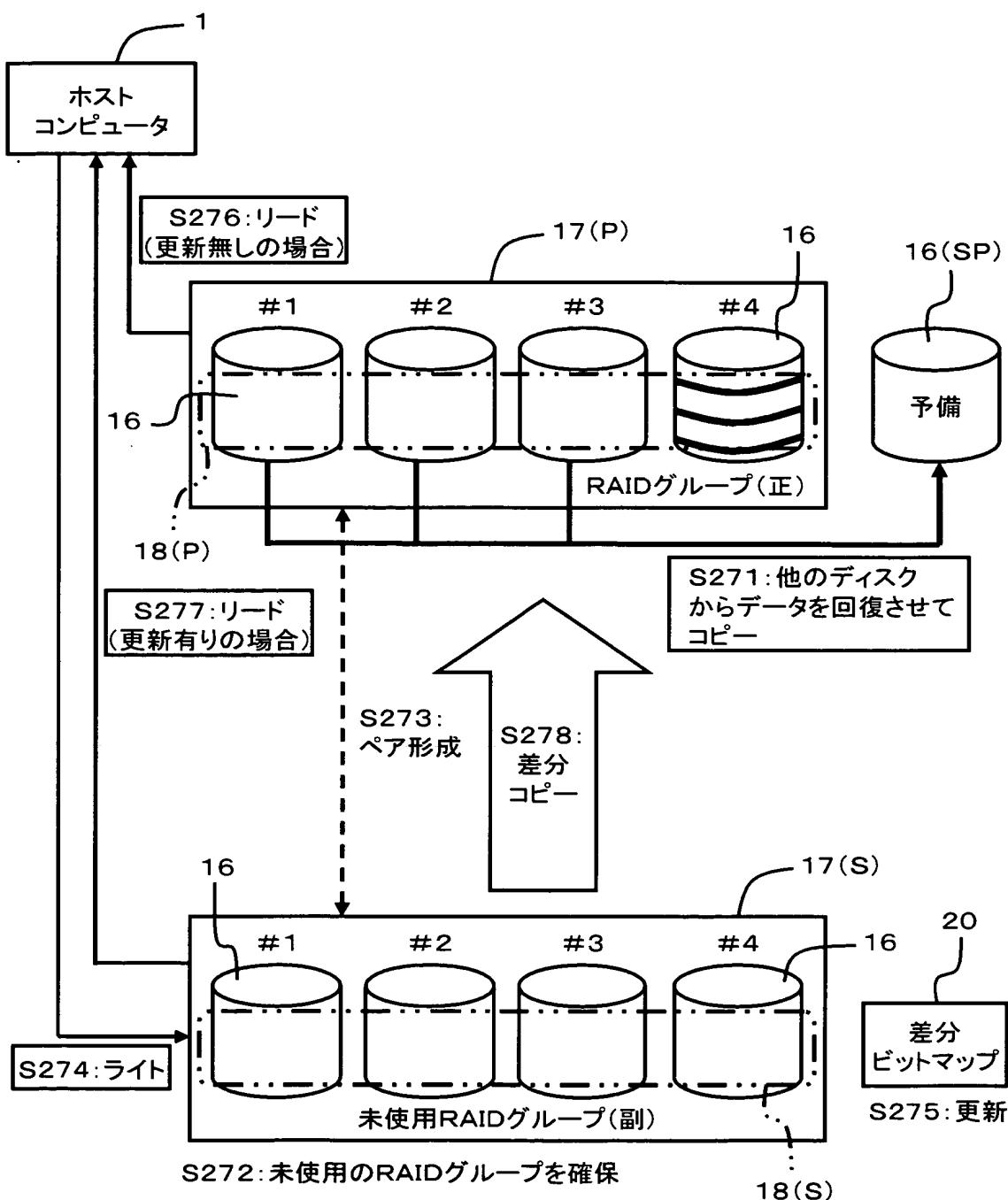
【図 23】



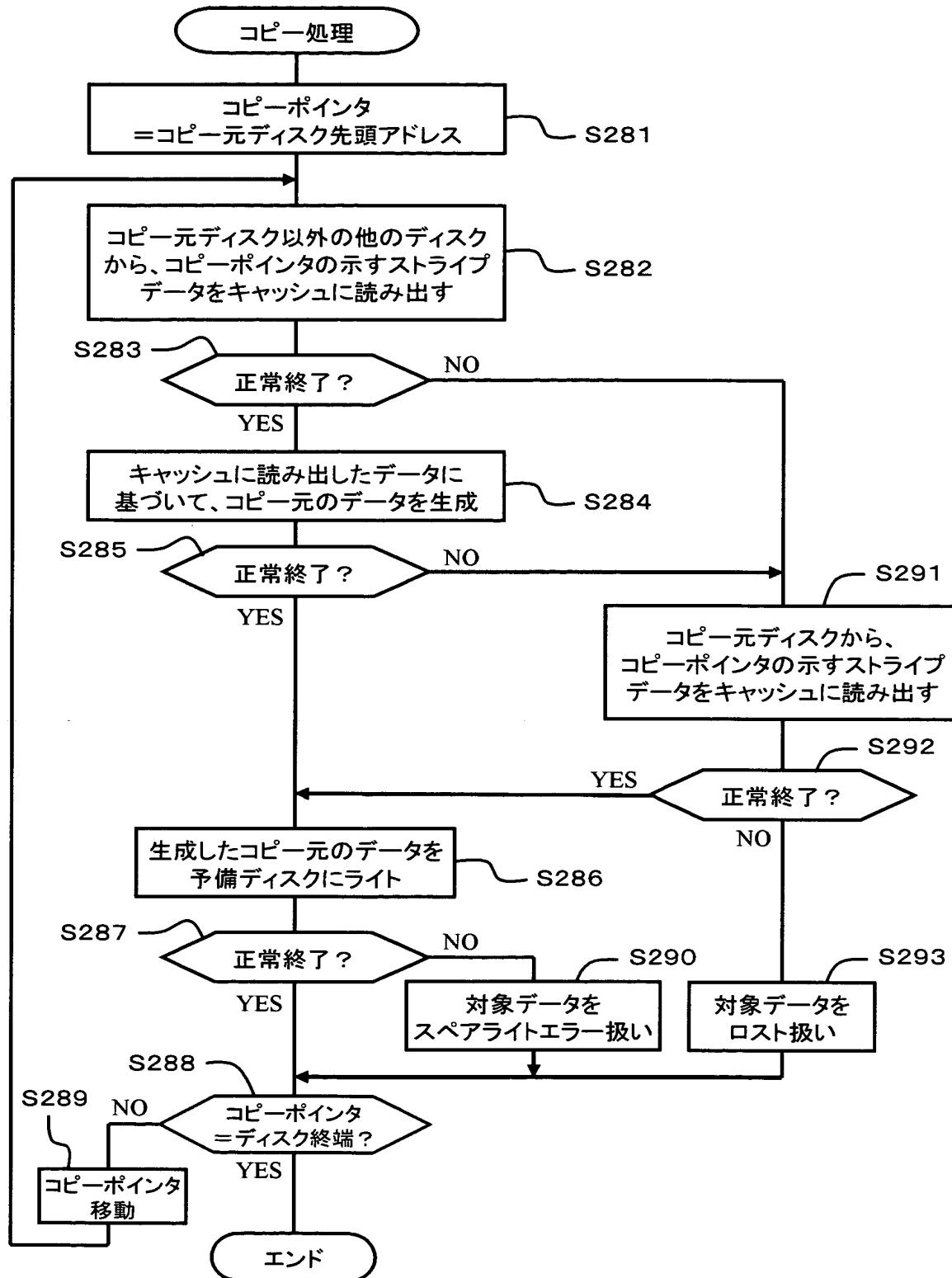
【図 24】



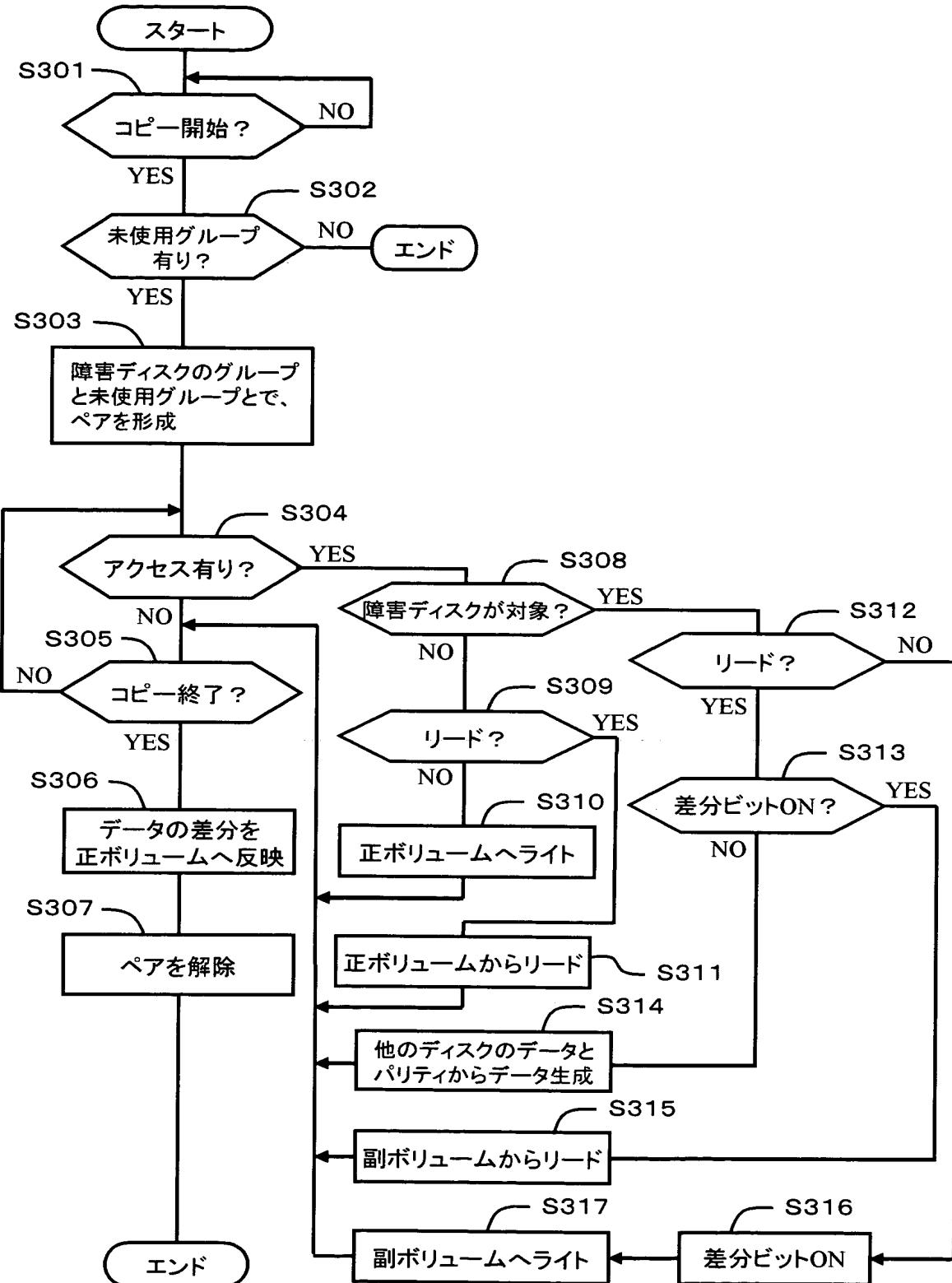
【図 25】



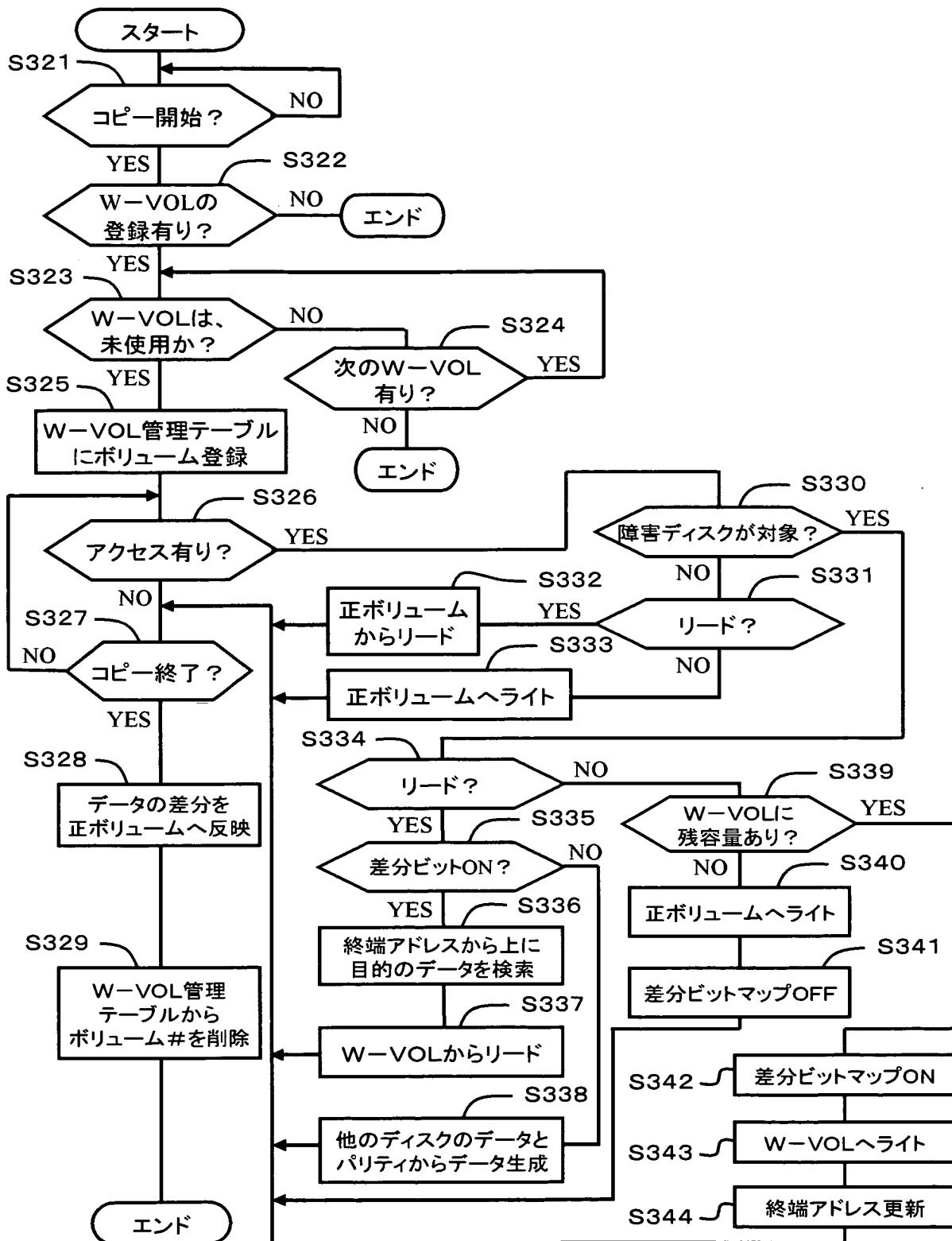
【図 26】



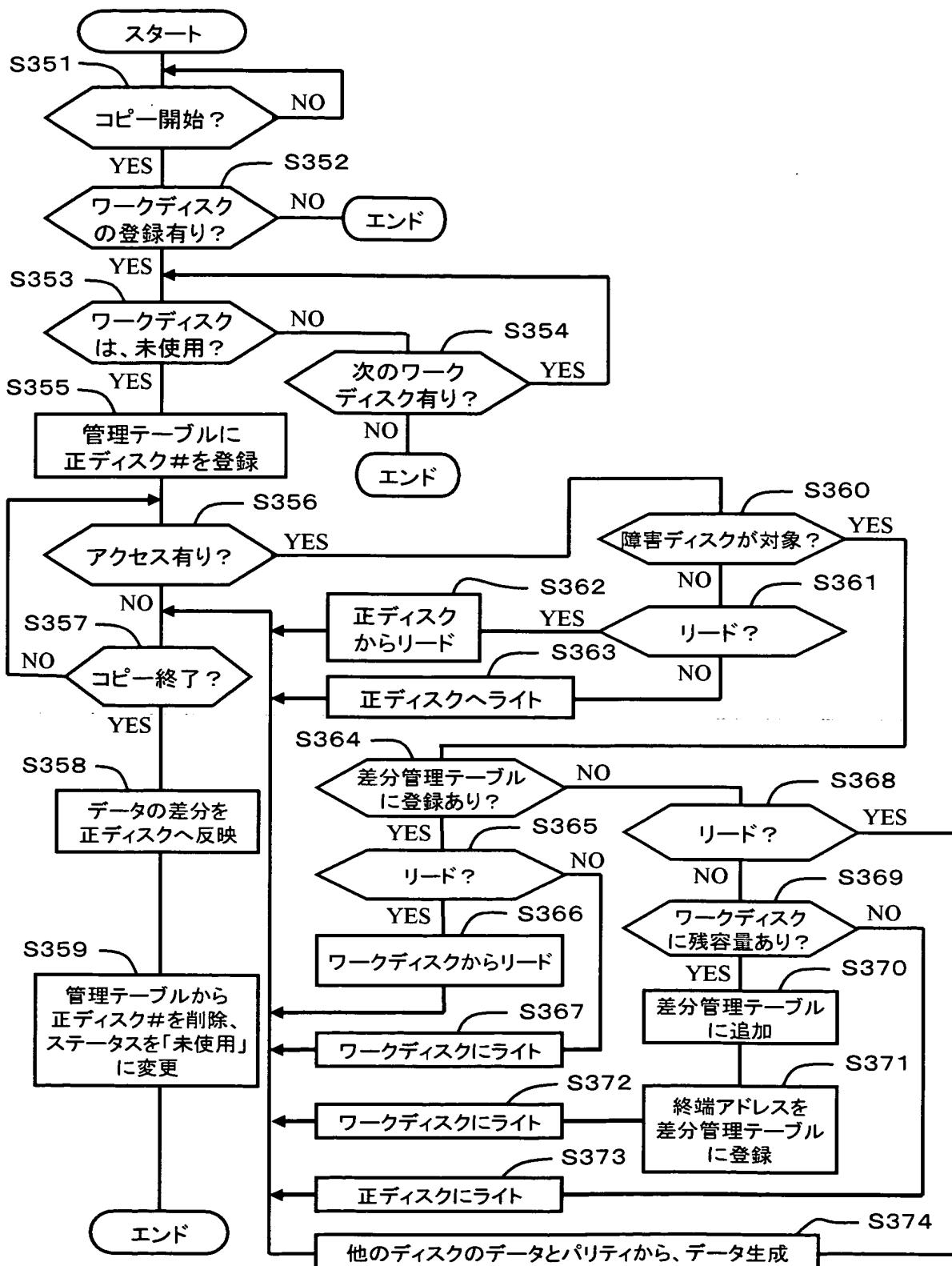
【図27】



【図 28】



【図 29】



【書類名】要約書

【要約】

【課題】 障害発生が予測されるディスクから予備ディスクへのデータ移行中に、ディスクアクセスを極力低減し、二重障害が引き起こされるのを未然に防止する。

【解決手段】 R A I D グループ17(P)を構成するディスク16(#4)に障害の発生が予測されると、ディスク16(#4)の記憶内容は予備ディスク16(SP)にコピーされる(S1)。コピーと同時にR A I D グループ17(P)とペアとなるR A I D グループ17(S)が設定され、副ボリューム18(S)が用意される(S2,S3)。書き込み要求は副ボリューム18(S)に対して行われる(S4)。差分ビットマップ20は、更新データを管理する(S5)。未更新データの読み出しは、正ボリュームから行われ(S6)、更新済データの読み出しは副ボリュームから行われる(S7)。データ移行が完了すると、副ボリュームの記憶内容が正ボリュームに反映される(S8)。

【選択図】 図5

認定・付加情報

特許出願の番号	特願2003-395322
受付番号	50301943726
書類名	特許願
担当官	第七担当上席 0096
作成日	平成15年11月27日

<認定情報・付加情報>

【提出日】	平成15年11月26日
-------	-------------

特願 2003-395322

出願人履歴情報

識別番号 [000005108]

1. 変更年月日 1990年 8月31日
[変更理由] 新規登録
住 所 東京都千代田区神田駿河台4丁目6番地
氏 名 株式会社日立製作所

2. 変更年月日 2004年 9月 8日
[変更理由] 住所変更
住 所 東京都千代田区丸の内一丁目6番6号
氏 名 株式会社日立製作所